DETERS, LAUREN B. F., Ph.D. Analysis of the Schizotypal Ambivalence Scale. (2017) Directed by Dr. John Willse, 108 pp.

The purpose of this study was threefold: a) to provide a thorough modern measurement example in a field where it is more limited in use, b) to investigate the psychometric properties of the Schizotypal Ambivalence Scale (SAS) through IRT measurement models, and c) to use the evaluation of the psychometric properties of the SAS to identify evidence for adherence to the relevant guidelines outlined in the *Standards for Educational and Psychological Testing* (hereafter *Standards*; AERA, APA, & NCME, 2014). Together, these goals were to contribute to the argument that the SAS is a robust measure of the ambivalence construct. An archived sample of over 7,000 undergraduate students was used to conduct all analyses.

Comparison of eigenvalue ratios indicated that the SAS data could be interpreted as essential unidimensional; however, results from the DIMTEST procedure (Stout, 2006) suggested a departure from unidimensionality. Results from the analysis provided adequate evidence for *Standard 1.13* (AERA, APA, & NCME, 2014).

The data were modeled via 1PL, 2PL, and 3PL models, and the 2PL model best fit the data. Examination of item-level statistics indicated that items 4, 8, 10, and 15 were endorsed more frequently than other items, and that items 2, 3, 9, 14, and 19 were the most discriminating. Items 7, 15, and 18 were flagged for possible misfit.

Results from the analysis of local independence revealed that many item pairs, particularly items 10 through 16, may have violated the assumption of local independence. Furthermore, results from DIMTEST (Stout, 2006) were significant, and

the partitioning of items indicated that items 12, 13, 14, and 16 were dimensionally similar. Analysis of the item-level characteristics provided evidence for *Standard 4.10* (AERA, APA, & NCME, 2014).

Analysis of information provided by the SAS at the item level revealed that items 2, 3, 9, 14, and 19 provided the most information at the higher end of the theta (θ) scale. Analysis of information at the scale level indicated that the reliability was $\alpha = .84$, and that the highest point on the TIF occurred at $\theta = 0.8$. The TIF and Cronbach's α provided evidence for *Standards 2.0* and *2.3*, and documentation of evidence for *Standard 7.12* (AERA, APA, & NCME, 2014).

Analysis of DIF revealed that only item 4 exhibited moderate DIF. In general, these results indicated that the SAS can be used across populations without concerns of items performing differently across subgroups. These results provided evidence for *Standard 4.10*, *Standard 3.0*, and *Standard 3.1* (AERA, APA, & NCME, 2014).

ANALYSIS OF THE SCHIZOTYPAL AMBIVALENCE SCALE

by

Lauren B. F. Deters

A Dissertation Submitted to
the Faculty of The Graduate School at
The University of North Carolina at Greensboro
in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Greensboro
2017

Approved by

_____
Committee Chair

For Chase.

APPROVAL PAGE

This dissertation written by LAUREN B. F. DETERS has been approved by the following committee of the Faculty of The Graduate School at The University of North Carolina at Greensboro.

Committee Chair _____

Committee Members _____

_____

_____

_____
Date of Acceptance by Committee

_____
Date of Final Oral Examination

ACKNOWLEDGEMENTS

Over the past several years, I have received support and encouragement from my department and external advisors. I would first like to thank my chair, Dr. John Willse, for assisting me from moving beyond my original concept to a cohesive study, and for his patience with helping me move forward as I managed my career and my dissertation. I would also like to thank my committee, composed of Dr. Tom Kwapil, Dr. Bob Henson, and Dr. Devdass Sunnassee, for their understanding, critical feedback, and positivity throughout this process. I would also like to thank my department for teaching me the lesson of communicating to a variety of stakeholders, a lesson that has served me well in both my career and my personal life.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER I

INTRODUCTION

Schizotypy represents the latent personality traits within people at-risk for developing schizophrenia (Meehl, 1962). The term "ambivalence," one of the latent traits found within schizotypy and schizophrenia, was originally coined by Bleuler (1911, 1950). Bleuler defined it as the simultaneous experience of both positive and negative emotions and the inability to integrate these emotions. He viewed ambivalence as a manifestation of thought disorder and he also considered it to be one of the core components of schizophrenia. Unlike accessory symptoms, he considered ambivalence to be a symptom everyone with schizophrenia experiences. Meehl (1962) initially proposed that ambivalence was a core symptom in schizotypy; however, he later amended this statement and said that ambivalence increased the likelihood of schizophrenia (Meehl, 1989). Despite its prominence in Bleuler's (1911, 1950) historical formulation of schizophrenia, as well as the emphasis Meehl (1989) placed on its role as a risk factor for developing schizophrenia, the concept of ambivalence has received relatively little attention in the mainstream psychopathology literature; specifically, the term did not appear in the DSM-IV-R under the entry for schizophrenia (Kuhn & Cahn, 2004), and does not appear in the DSM-V (American Psychiatric Association, 2013) under the entry for schizophrenia. Additionally, little work has been done to uncover the ambivalence construct as defined within schizotypy and schizophrenia.

However, one measure of ambivalence has been attempted. Raulin (1984) developed the Intense Ambivalence Scale (IAS) with the purpose of measuring the ambivalence construct as defined under schizotypy. However, after using the scale in studies, he later revised the scale to include fewer items, and this revised scale is now known as the Schizotypal Ambivalence Scale (SAS; Raulin, 1986). The SAS is intended to measure the ambivalence construct defined under schizotypy and schizophrenia. Though the SAS shows promise, the scale has not been through rigorous psychometric analysis. And, though a great deal of importance has been placed on the role of ambivalence in schizotypy and schizophrenia, the construct has not been tested with more modern measurement theories such as item response theory (IRT). Thus, the study aimed not only to investigate the ambivalence construct through evaluating the psychometric properties and construct validity of the SAS, but to subject the SAS to modern measurement theories.

A Brief History of Psychological Measurement

In the variety of psychological fields, factor analysis, including both confirmatory factor analysis (CFA) and exploratory factor analysis (EFA), and structural equation modeling (SEM) have proven to be the more popular methods for establishing, evaluating, or refining measurements. From 1993 to 1997, MaCallum and Austin (2000) found nearly 500 psychological journal articles with applications of SEM. Furthermore, fields in psychology have heavily used SEM for both cross-sectional and longitudinal studies (MaCallum and Austin, 2000). The popularity of SEM and CFA methods has been further propagated by the creation of user-friendly software packages.

Classical test theory (CTT; Lord, 1980) has also seen application in psychological testing for decades. However, to create scales that meet more rigorous assumptions, psychologists have been slowly moving toward IRT, or a combination of CTT and IRT. Embretson (1996) first discussed CTT and IRT for psychologists, specifically regarding the advantages that IRT permits. Though IRT was originally introduced in the 1950s and 1960s by Lord (1952), its use beyond the field of education has been slow. In her conclusions, Embretson (1996) hypothesized several reasons for psychologists' hesitance to embrace IRT. In addition to the idea that IRT requires more statistical knowledge than CTT or CFA, her reasons included that at the time of her article, there was no IRT textbook for psychologists (Embretson and Reise went on to write such a book in 2000). Indeed, she conceded that "there is no such book available, probably primarily because IRT experts have been more concerned with technical issues than with expositional issues" (Embretson, 1996, p. 348). And, despite the prevalence of psychometric software currently available, most software packages are specialized and thus not provided within typical statistical packages used in psychology. However, the field has seen recent growth in the use of IRT.

Since the publication of Embretson's (1996) article, several fields in psychology have begun to embrace IRT. For example, IRT has been used to establish and analyze scales such as the Mindful Attention Awareness Scale (Van Dam, Earlywine, & Borders, 2010), the Hypomanic Personality Scale (Meads & Bentall, 2008), the Positive and Negative Syndrome Scales (Santor, Ascher-Svanum, Lindenmayer, & Obenchain, 2007). IRT has also been used to evaluate the Posttraumatic Stress Disorder (PTSD) Checklist

(King, Street, Gradus, Vogt, & Resick, 2013). Others have gone beyond the typical IRT models and have applied the graded response model (GRM; Samejima, 1969). Most recently, Zanon, Hutz, Yoo, and Hambleton (2016) illustrated an application of the GRM with the Affect Scale, a scale intended to measure a person's positive and negative affect.

While the expansion of use in the field of psychology is promising, IRT is not use at a rate comparable with CFA or CTT. Given the shortcomings of CTT and the advantages of IRT, discussed further in the review of literature, as well as the multitude of IRT models from which one can choose, the use of IRT in psychology would strengthen scales. IRT should not be used in place of CTT – the two theories are complementary – but the use of IRT is necessary for a field that makes diagnostic decisions based on an individual's responses to a psychological assessment.

Standards for Educational and Psychological Testing

At the beginning of assessment development, educational and psychological researchers should have a clear, organized process by which they develop an assessment for evaluating a construct. Further, these assessments should be subject to a thorough evaluation, either externally or internally, and should provide clear guidance on interpretability of results. To provide a set of standardized rules which researchers in both education and psychology can follow, three organizations collaborated to revise the *Standards for Educational and Psychological Testing* (hereafter *Standards*; AERA, APA, & NCME, 2014). The *Standards* provide uniform guidelines in education and psychology for evaluating an assessment's validity, reliability, and fairness, as well as guidelines for the development process of any assessment. Additionally, the *Standards* provide

guidance on interpretability and ethical use of results, and detail the use of the *Standards* for credentialing/licensing programs, psychological assessments, and accountability systems.

While the *Standards* (AERA, APA, & NCME, 2014) are not legal guidelines to which an organization must adhere, they reflect "the best professional judgments about how to design, develop, and use tests that provide scores with high levels of validity, reliability, and fairness" (Plake & Wise, 2014, p.10). Previous iterations of the *Standards* have been referenced as an argument in court cases examining test use in high-stakes assessments, highlighting the importance of using the *Standards* in adhering to best practices (Plake & Wise, 2014, p. 10). And, though the *Standards* are primarily used in formal assessment development settings, they outline guidelines that anyone developing an assessment should follow, including those scales or surveys used diagnostically or to predict risk of later issues. Notably, the revised *Standards* continue to exclude references to less formal assessments, such as interim assessments in the classroom used for formative purposes, though the joint committee for developing the *Standards* reported that "classroom teachers would benefit from reading the *Standards*" (Plake & Wise, 2014, p. 6). In summary, the *Standards* provide a universal set of rules that ensure that assessments – especially those used to make decisions about students' placement or those used to classify disorders – are valid, reliable, and fair.

Purpose and Relevance of the Research

The purpose of this study was threefold: a) to provide a thorough modern measurement example in a field where it is more limited in use, b) to investigate the

psychometric properties of the SAS through three measurement models, and c) to use the evaluation of the psychometric properties of the SAS to identify evidence for adherence to the relevant guidelines outlined in the *Standards* (AERA, APA, & NCME, 2014), thus providing a beginning argument for the validity and reliability of the SAS. The SAS was evaluated at both the item and scale levels by using a combination of classical test theory (CTT) and IRT methods, as well as differential item functioning (DIF). Results from the analyses were provided as evidence for the relevant *Standards* (AERA, APA, & NCME, 2014).

IRT is foremost used in educational testing. However, researchers have begun to use IRT in assessing psychological scales. Notably, IRT has been used to establish other schizophrenia psychological assessments such as the Wisconsin Schizotypy Scales (Winsterstein, Ackerman, Silvia, & Kwapil, 2011) and Interview Report Scales of Cognitive Impairment (Reise et al., 2011). However, IRT studies in psychology are often only one of a few of their kind for each subject area.

Relatively little research has been done on the construct of ambivalence, even though well-respected psychologists have endorsed the concept of ambivalence as an integral part of schizotypy and schizophrenia. A study thoroughly investigating the psychometric properties of the SAS, the only known scale for measuring ambivalence, would provide more confidence in the use of the SAS as a measure of the ambivalence construct.

Such a study was essential to addressing the gaps in literature regarding the measurement of ambivalence. Though not a widely-used scale, the SAS is a promising

measure for examining levels of ambivalence in individuals. If ambivalence in schizotypic individuals is associated with a higher risk for developing full-blown schizophrenia as Meehl (1989) suggested, then analyzing the psychometric properties of the SAS is beneficial to those conducting schizophrenia research and those at-risk for developing schizophrenia. High scorers on the IAS and SAS included those who exhibited schizotypal symptoms and those who were schizophrenic patients, but the scale also captured those with unrelated conditions. If the SAS is analyzed with modern measurement theories, problem items can be targeted via point biserial statistics, IRT estimated *a*-parameters, and DIF. Additionally, researchers can refine how they operationalize the construct of ambivalence, ultimately resulting in a more accurate measure of ambivalence as defined within schizotypy and schizophrenia. This would provide additional insight to researchers and clinicians looking to examine individuals at-risk for schizophrenia and identify the predisposition for schizophrenia earlier in life.

This study goes one step further in closing the gap between the research methodologies used in the fields of psychology and educational research methodology. Item response theory (IRT) is most frequently used in educational test development, and most often the capabilities of IRT are shown in the context of education, specifically achievement testing. Currently, most of the scales used in any field of psychology have been assessed with CTT, the limitations of which are outlined above (Embretson & Reise, 2000). Evaluating the quality of the SAS through multiple IRT models provides an opportunity to conduct a more demanding analysis of the SAS, thus integrating analyses used in educational research methodology with psychological assessments. Analysis of

the SAS with IRT provides an alternative use of the measurement models most notably used in educational testing, highlighting the robust uses of such models. Ultimately, the use of IRT analyzing the psychometric properties of the SAS was helpful in a) emphasizing the diverse fields to which it can be applied and highlighting many of its strengths, and b) providing details for improvement in the SAS.

Finally, the purpose of the study was to the gaps between the fields of psychology and education by connecting results of analyses to the *Standards* (AERA, APA, & NCME, 2014). The fields share a common set of guidelines for developing assessments, evaluating assessments, and interpreting scores; by using the *Standards* to establish evidence for validity, reliability, and fairness of the SAS, the study connects the relevant *Standards* to the measurement approaches used here and more clearly connects the two fields.

First, the concepts of schizotypy and schizophrenia are introduced. The review of the literature begins with a discussion of the concept of ambivalence as defined under schizophrenia, as well as the historical literature on the attempts made to measure ambivalence. The literature on construct validation of ambivalence are discussed, including studies aimed at validating constructs using modern measurement theories. The discussion then shifts to measurement theories, most prominently IRT. This component includes an in-depth review of the major properties and assumptions under IRT. The purpose and relevance of the study within the context of two fields – educational measurement and psychology – is introduced, followed by the focal research questions. The Methodology chapter introduces the procedures and their relevance in the context of

each question, and clearly connects each analysis to the relevant *Standards* (AERA, APA, & NCME, 2014). The Results chapter details the results of the analyses as they pertain to each research question. Finally, the Discussion chapter summarizes findings from each of the research questions, areas of the SAS in need of further research, and implications of findings as they relate to the *Standards*.

CHAPTER II

REVIEW OF THE LITERATURE

The literature in this chapter is divided into three sections: schizophrenia and ambivalence, construct validity, and measurement theory. The first part of the chapter introduces ambivalence in the context of schizophrenia, and then discusses the background on ambivalence and the development of the construct; then, the limited instances in which the construct of ambivalence has been measured are discussed. The second part of the chapter addresses the concept of construct validity and the importance of construct validity as it relates to dimensionality. Lastly, modern measurement theory, the major assumptions of IRT, the background on models, and DIF are discussed. The relevant *Standards* (AERA, APA, & NCME, 2014) are referenced throughout.

Schizophrenia

Schizophrenia is one of the most severe psychopathological illnesses, and one of the most difficult to predict (Lenzenweger, 2006). It currently affects one in 100 individuals. Although the exact risk factors are not clear; a large portion of research has been dedicated to examining those who are at-risk for developing schizophrenia but ultimately who do not develop schizophrenia (Lenzenweger, 2006). Schizotypes are those who are at-risk for developing schizophrenia, a large part of which includes genetic predisposition. Per Lenzenweger (2006), schizotypes "possess … schizotypy, or a latent

ability personality organization that harbors the genetic liability for schizophrenia" (p. 163).

Schizophrenia can be markedly different from one individual to the next. The DSM-V (APA, 2013) includes five different types of schizophrenia. However, the DSM-V (2013) specifies that three diagnostic criteria must be present to be diagnosed with schizophrenia: at least two characteristic symptoms, which include hallucinations, disorganized thought, delusions, disorganized behavior, and negative symptoms (such as blunted affect); social or occupation dysfunction; and duration (symptoms must persist for at least six months). Historically, Bleuler (1911, 1950) described four core symptoms (known as the four a's) all schizophrenics will exhibit: loosening associations, disturbances in affectivity, autism, and ambivalence, the last of which is the focus of the SAS.

Ambivalence

Bleuler first alluded to ambivalence in 1904 in an article covering negative suggestibility (Kuhn & Cahn, 2004). Though the term and construct of ambivalence was not yet coined, he described it as two opposing forces. Shortly after his first mention of ambivalence, Bleuler (1911, 1950) discussed the concept of ambivalence and its role in the schizophrenia disorders, specifically describing ambivalence as one of the four central components of schizophrenia. For Bleuler (1911, 1950), ambivalence occurred when "contradictory feelings or thoughts exist side by side without influencing each other" (pp. 354-355). Bleuler (1911, 1950) believed there were three types of ambivalence: voluntary ambivalence, intellectual ambivalence, and emotional ambivalence. Voluntary

ambivalence is the conscious disagreement over doing something or not doing something; intellectual ambivalence refers to ambivalence in thinking and reasoning; and emotional ambivalence refers to polarizing feelings directed at the same person or object simultaneously. In all cases, however, ambivalence was characterized by the difficulty or inability to integrate disparate thoughts and feelings. Bleuler (1911, 1950) viewed this as an example of schizophrenic thought disorder.

In describing his theory on the risk factors for schizophrenia, Meehl (1962) also discussed ambivalence. Though he initially agreed with Bleuler's role of ambivalence as a core component in schizophrenia, he later revised his theory (Meehl, 1989). Notably, he commented that while many people are genetically predisposed to schizophrenia, not all go on to develop schizophrenia. He theorized that those people who possessed the ambivalence construct would be more likely to develop full-blown schizophrenia, noting that ambivalence was a risk factor for schizophrenia.

The ambivalence construct has been recognized in a variety of psychological disorders such as schizophrenia and depression (Sincoff, 1990), but there does not exist unanimous agreement in how to operationalize the term. Sincoff (1990) examined the definition of ambivalence in a variety of psychological fields and attempted to understand how ambivalence could be measured. Even though the ambivalence construct was prominent in early theories of schizophrenia, the construct has been most widely employed in the psychoanalytic literature. Freud referenced ambivalence in both 1913 and 1917 in discussions on the ego and death (Freud, 1913; Freud, 1917), in which he defined ambivalence as "the sway of contrary tendencies," and referenced the emotional

ambivalence that Bleuler discussed (Freud, 1913, p. 60). Though other psychoanalysts have discussed the construct of ambivalence, only Freud's definition approached what Bleuler believed (Sincoff, 1990). Sincoff (1990) also noted that other theorists have identified ambivalence as an inherent piece of obsessive-compulsive personalities and depression, and sociologists have also described ambivalence as a social issue. Ambivalence is also recognized in disorders related to but markedly different from schizophrenia, such as borderline personality disorder (Kernberg, 1984). Furthermore, other definitions of ambivalence are measured, such as through the Ambivalence Over Emotional Expressiveness Questionnaire (AEQ; King & Emmons, 1990), which intends to tap into an emotional ambivalence construct. It is common in psychology for "different diagnostic systems [to] use the same diagnostic term to describe different systems" (AERA, APA, & NCME, 2014, p. 160). However, the prevalence of ambivalence across fields and measures in the field of psychology suggests that the operational definition of ambivalence depends on how it is defined in the context of the disorder.

Measurement of Ambivalence

The literature on the concept of ambivalence is limited despite the prominence of the construct in relation to schizotypy and schizophrenia, as relatively few articles and chapters have been written on the subject (Raulin & Brenner, 1993; Kwapil, Mann, & Raulin, 2002). Furthermore, few recent studies have focused on operationally defining the ambivalence construct, despite its early prominence in the literature. In 1984, Raulin sought to develop a measure specifically aimed at measuring the ambivalence construct (Raulin, 1984). Raulin (1984) created the Intense Ambivalence Scale (IAS), a 45-item

true/false measure with the purpose of identifying ambivalence that characterized risk for developing schizophrenia. In creating his scale, Raulin (1984) administered the scale to college students through several iterations, altering and removing items with each administration. He then interviewed 72 of the college students that completed the scale for validating the measure. To further evaluate the scale, Raulin (1984) administered the IAS to patients with schizophrenia and patients with depressive disorders, psychology clinic outpatients, and a group of control subjects. Raulin (1984) found that students who scored higher on the scale more often reported feelings of ambivalence during the interviews, and that schizophrenic patients scored higher on the measure than the control group. However, schizophrenic patients did not significantly differ from psychology outpatients, and depressed patients outscored schizophrenic patients on the scale.

Kwapil, Raulin, and Midthun (2000) further examined the IAS in a 10-year longitudinal study with the purpose of examining the predictive validity of the IAS. Participants were chosen from a longitudinal study examining psychosis proneness (Chapman, Chapman, & Kwapil, 1994), and participants in this study received a variety of measures of schizotypy. Of the sample, 203 subjects were chosen because they received a standard score of 1.96 on the Perceptual Aberration Scale and the Magical Ideation Scale. The control group consisted of 159 participants who did not receive a standard score higher than 0.50. The initial contact with participants involved a diagnostic interview and mass screening. The 10-year follow-up included an interview. Kwapil et al. (2000) found that high scores on the IAS at the initial assessment were associated with subsequent development of schizotypal symptoms, drug abuse, and

alcohol abuse. However, high scores on the IAS were also related to psychoses other than schizophrenia. The scale also predicted psychotic-like symptoms and symptoms of depression 10 years later. Given that high scores on the IAS were associated with more than the subsequent development of schizotypal symptoms, it appeared that the IAS was capturing more than the construct of ambivalence as it relates to the schizophrenias.

Given the mixed results from the original and later administrations of the IAS, specifically, the concern that the IAS measured one's risk for depression more than one's risk for schizophrenia, Raulin (1986) created the Schizotypal Ambivalence Scale (SAS), a 19-item measure designed to examine ambivalence as defined under schizotypy (see Appendix A). Specifically, Raulin (1986) retained 12 items from the IAS and developed seven new items for the SAS. The intent was to create a scale that better delineated between ambivalence as defined under schizotypy and ambivalence as defined under depression and other forms of psychopathology.

Kwapil, Mann, and Raulin (2002) examined the psychometric properties of the SAS. Specifically, they examined ethnicity and race differences in the SAS, as well as the reliability and concurrent validity of the SAS. In the study, 997 college students were given the SAS during a mass screening of multiple schizophrenia-related measures; 131 of those participants were also administered a diagnostic interview. Kwapil et al. (2002) found that the reliability of the SAS was comparable to the reliability of the IAS – which was 26 items longer – and that it correlated moderately with other measures of schizotypy. They also found that there were no significant differences for ethnicity or gender in the total score. The SAS accounted for variation in schizotypic symptoms

above and beyond other questionnaire measures of schizotypy. And, unlike the IAS, high scores on the SAS were not associated with substance abuse or major depressive symptoms. Kwapil et al. (2002) also found that high scores on the SAS were related to schizotypal, schizoid, and paranoid symptoms. Ultimately, Kwapil et al. (2002) demonstrated that the SAS better captured the ambivalence construct as defined under schizophrenia, whereas the IAS appeared to assess a broader conceptualization of ambivalence shared by depression and other forms of psychopathology.

Mann, Vaughn, Barrantes-Vidal, Raulin, and Kwapil (2008) built upon the work conducted by Kwapil et al. (2002) by examining the test-retest reliability of the SAS, as well as the concurrent validity in a sample of high scorers. The SAS was administered to 1,798 college students as part of a mass screening; of those participants, 122 volunteered to be retested eight to 11 weeks after the mass screening. Another subset of participants was administered a diagnostic interview. Like the results found by Kwapil et al. (2002), there were no gender differences across scores on the SAS. However, Mann et al. (2008) found that African Americans tended to score higher than Caucasians. Overall, Mann et al. (2008) found the overall reliability and the test-retest reliability of the SAS was high. The measure correlated moderately with other measures of schizophrenia symptoms, and Mann et al. (2008) found that those who scored high on the SAS had higher interview ratings of schizotypal, schizoid, paranoid, psychotic-like, and negative symptoms than the control group. The researchers concluded the SAS to be a promising measure of ambivalence as defined under schizotypy.

Structure of the schizotypal ambivalence scale. Though a few studies have demonstrated that the SAS is a reliable measure of ambivalence as defined under schizotypy, fewer studies have investigated the internal structure, or dimensionality, of the SAS. MacAulay, Brown, Minor, and Cohen (2014) sought to investigate the structural properties of the SAS. The SAS, along with other questionnaires, was administered to 334 participants. However, the SAS was administered in Likert format, in which participants could select from 1 (strongly disagree) to 5 (strongly agree). MacAulay et al. (2014) conducted an EFA using PCA with an oblique rotation and found three underlying factors: "interpersonal ambivalence, indecision/insecurity, and contradictory feelings" (p. 796). The authors further proposed that the factors "seemed to closely resemble Bleuler's multifactor conceptualization of ambivalence" (p. 796). Despite the use of Likert-style items rather than dichotomous items, these results and the absence of other studies warrant further investigation into the SAS.

Though only a few studies conducted an in-depth analysis of the SAS, results have shown that the SAS is a reliable scale for measuring ambivalence that is characteristic of schizotypy and schizophrenia. It also correlates moderately with other scales designed to assess schizotypy. However, the existence of ambivalence among other disorders other than schizophrenia and the correlations of the scale with other measures, especially those measures not related to schizotypy or schizophrenia, highlight the need for additional analysis of the operational definition of ambivalence under schizophrenia, and of the SAS as a tool to measure that operational definition. Furthermore, because "diagnostic criteria may vary from one nomenclature system to

another" (AERA, APA, & NCME, 2014, p. 160), it is important to explicitly define the criteria and the latent trait.

Literature, or the lack thereof, has illustrated the difficulty in defining and measuring the ambivalence construct. Sincoff (1990) suggested that issues regarding whether ambivalence is a trait or something that arises in certain contexts, and well as whether ambivalence can even be measured with self-reports, should be considered. Despite the acknowledgement by Bleuler (1911, 1950) and Meehl (1962, 1989) of ambivalence as an integral part of schizotypy, little research has been done on the ambivalence construct beyond the development of the IAS and SAS, suggesting a gap in the literature and need for such research.

Construct Validity

Establishing construct validity is an important part of evaluating psychological measures, but requires an understanding of the term "construct." Most often latent, a construct is a "postulated attribute of the people, assumed to be reflected in test performance" (Cronbach & Meehl, 1955, p. 283). Cronbach and Meehl (1955) also argued that the construct must occur within a nomological net, or within a set of laws, some of which are observable.

However, Kane (2012) has argued that while Cronbach and Meehl (1955) laid the foundations of what we now interpret as construct validity, "their definition of 'construct,' which required a strong theory or nomological net, has not stood the test of time" (p. 68). He argues that the construct must be well-thought out using interpretive arguments (IA) – as established by Kane (2006) – so one avoids the drawbacks of

evaluating constructs that have arisen from poorly constructed theories and the liberal use of the term "construct." Messick (1988) further cautioned against the misuse of test scores, specifically the interpretation of scores, as these incorrect interpretations "can often be traced to construct under-representation or construct-irrelevant variance" (Slocum-Gori & Zumbo, 2011, p. 444). Thus, the validity field has moved away from solely defining separate types of validity and moved towards a model of validity that involves constructing arguments and evaluating validity within and across a system.

The SAS is a measure of the ambivalence construct as defined under schizotypy and schizophrenia and established through years of research. Those who possess more of the ambivalence construct than others should have higher scores. However, both the IAS and SAS demonstrated that impairments experienced by participants with high scores were not limited to schizotypic symptoms. Even though the construct is not new, there is a not a clear distinction of the varying definitions of ambivalence across the different disorders. As noted several times already, the ambivalence construct is present in fields other than schizophrenia, and the ambivalence construct is not well understood (Sincoff, 1990). Due to the prevalence of ambivalence across disorders distinctly different from schizotypy schizophrenia, establishing the construct is essential to effectively measuring ambivalence in those prone to schizotypy. Is the construct of ambivalence operationally different across fields? Is ambivalence as measured by the SAS unidimensional? It was not the purpose of this study to establish construct validity of the SAS or construct the interpretive argument for ambivalence throughout this paper. Instead, establishing

dimensionality is just one "essential component of construct validity" (Slocum-Gori & Zumbo, 2011, p. 443), Indeed, the *Standards* (AERA, APA, & NCME, 2014) clarify

> Analyses of the internal structure of a test can indicate the degree to which the relationships among test items and test components confirm to the construct on which the proposed test score interpretations are based … The extent to which item interrelationships bear out the presumptions of the [conceptual] framework would be relevant to validity. (p. 16)

The argument is that that one must first establish the dimensionality of the SAS to fully understand the data that result from it and to be able to verify its construct validity. To ensure the interpretability of scores that result from the SAS, it is necessary to first establish the dimensionality of the SAS: "One way, among many, to prevent inappropriate consequences from test score interpretation is to focus on the theoretical dimensions of the construct a test is intending to measure" (Slocum-Gori & Zumbo, 2011, p. 444).

Dimensionality

Before evaluating model fit, one must first establish that the construct being measured by a scale is unidimensional, as a unidimensional model will not adequately capture the characteristics of multidimensional data. The use of a unidimensional model on multidimensional data can lead to misuse of scores or errors in interpretation, both of which have consequences in educational and psychological testing. For example, if a scale used to diagnose depression has items that are also capturing anxiety, which has a high comorbidity rate with depression, then diagnoses derived from the use of the scale may be a misrepresentation of the underlying issue within individuals. *Standard 1.13* in

the *Standards* explicitly states that evidence of the relationship of items should be provided if the interpretability of the evidence "depends on the premise about the relationships among test items" (AERA, APA, & NCME, 2014, pp. 26-27). Therefore, one must establish dimensionality to defend the interpretability and use of test scores.

"A fundamental assumption of test theory is that a score can only have meaning if the set of items measures only one attribute or dimension" (Hattie, Krakowski, Rogers, & Swaminathan, 1996, p. 1). Across varying fields, researchers usually work under the assumption that a scale or a test is intended to measure one construct, or one dimension. That is, the items that comprise a test or scale are all intended to measure one latent construct, thus confirming that the test or scale is unidimensional. This assumption of unidimensionality is ideally met during the original construction of the test or scale, where existing literature is used to create a set of items to measure a latent trait. However, unidimensionality is often undermined by construct-irrelevant factors. In educational testing, these "might include level of motivation, test anxiety, ability to work quickly… in addition to the dominant [factor] measured by the set of items" (Hambleton & Swaminathan, 1985, p. 17). Given the presence of ambivalence across psychological fields and its varying definitions, the first step in evaluating the SAS is assessing its dimensionality.

While unidimensionality of a scale is usually a researcher's goal – and the assumption of modern measurement theories – it is unreasonable to assume that the items assess only one common factor or dimension and excludes all possible extraneous yet related factors, as discussed above. There will always be construct-irrelevant factors that

interfere with the factor of interest, and despite adherence to best practices and existing research, some variation in scales will nearly always be attributed to something other than the factor of interest. In the development of mathematics assessments for example, there is a continuous debate over minimizing the need for students to possess language skills to successfully answer a mathematical question in vignette form. One might postulate that there is always some measure of language skills on these assessments, regardless of how small its contribution is to the overall assessment.

Thus, analyses of dimensionality often yield results that do not truly reflect one dimension. Slocum-Gori, Zumbo, Michalos, and Diener (2009) established that while many psycho-educational scales purport to be strictly unidimensional, most are essential unidimensional. Essential unidimensionality reflects the notion that there are minor latent traits or factors in addition to one dominant factor: "strict unidimensionality… is defined as one dominant latent trait variable with no secondary minor dimensions" (Slocum-Gori & Zumbo, 2011, p. 446).

Major concepts in unidimensionality. Literature tells us that there are multiple methods and indices by which one can evaluate dimensionality of a scale, and that some are more robust measures of dimensionality than others (Hattie, 1985). One of the earlier methods of establishing unidimensionality was through analyzing answer patterns, a method that preceded reliability indices. In Hattie's (1985) thorough review of methods to assess unidimensionality, he described these indices as "ideal scale pattern occurs when a total test score equal to *n* is composed of correct answers to the *n* easiest questions, and thereafter of correct answers to no other questions" (p.140). Guttman

(1944) and Green (1956) developed indices based on answer patterns, though all three methods pose issues. Specifically, these indices are based on the notion of reproducibility, but "perfect reproducibility may not necessarily imply that [items on a scale] are unidimensional" (Hattie, 1985, p. 143).

Developed around the same time as answer pattern indices, reliability indices appeared in the field. One of the most common indices, is Cronbach's α (Cronbach, 1951). While Cronbach's α has stood the test of time as a reliability index, Davenport, Davison, Liou, and Love (2015) argue that using Cronbach's α as an index of unidimensionality is confounding the properties of reliability, dimensionality, and internal consistency. They argue that even though these three properties are related, they should be established through separate indices. Other reliability indices were established, such as mean correlations (Cronbach, 1951), but the use of this method to establish unidimensionality still confounds reliability, dimensionality, and internal consistency.

Another method established to test unidimensionality is the reduction of data and extraction of factors, known as factor analysis or principal components analysis. Principal components analysis (PCA), which is related to both EFA and CFA, transforms data into a set of components or factors that are not correlated. The items on a scale load onto one of the components or factors, accounting for variance in the data. The first component extracted will always account for the most variance, and this percent of variance has been used as an index for unidimensionality (Hattie, 1985). The main theory is that the larger the amount of variance the first component comprises, the more likely the scale is unidimensional.

Others have used the eigenvalues of the components as indicators of unidimensionality. For example, Kaiser (1960) argued that only eigenvalues that are greater than one should be counted towards dimensionality. Another principal component approach uses the ratio of eigenvalues to establish essential unidimensionality. One of the more common but broadly debunked rules involves counting only factors with eigenvalues greater than one (Kaiser, 1960). If only one eigenvalue is greater than one, then a test is unidimensional. However, many in the field are critical of this method, going so far to say that "too many researchers blindly rely on the so-called eigenvalue-greater-than-one rule" (Thompson & Daniel, 1996, p. 200). Furthermore, others have found that using this method typically results in an overestimation of the number of actual factors (Zwick & Velicer, 1986).

Other rules involving eigenvalues have seen greater success, such as the examination of eigenvalue ratios. Specifically, if the ratio of the first to second eigenvalue is approximately 4:1 (Lord, 1980) or 3:1 (Slocum-Gori & Zumbo, 2011), one has evidence of essential unidimensionality. Hattie (1985) argued that this approach could fail in cases where the difference between the second and third eigenvalues is large. However, Slocum-Gori and Zumbo (2011) compared multiple criteria from factor analysis approaches for assessing strict and essential unidimensionality with five conditions varied (including sample size) and found that the eigenvalue ratio rules of 4:1 and 3:1 – were suitable for assessing essential unidimensionality.

Related to PCA is factor analysis, specifically CFA and EFA. Often used in psychology to evaluate an existing measurement, factor analysis can be also used to

reduce data produced from a scale to a common latent trait, confirm the underlying structure of the data produced from a scale, and to explain the interrelationships of the items on the scale. It differs from PCA because factor analysis "estimates a uniqueness for each item given a hypothesis as to the number of factors" (Hattie, 1985, p. 146). And, unlike PCA, existing research is the main driver of interpreting factors, not statistical methods. Despite some of its limitations in analyzing dichotomous data, there are several methods employed to diminish this limitation such as the use of tetrachoric correlations rather than Pearson product-moment correlations. Factor analysis continues to be among the more popular methods of assessing dimensionality. Researchers conduct CFA when the underlying structure of the data is known – usually based on experience or existing research. Exploratory factor analysis (EFA) is used when researchers do not know the underlying data structure or the interrelationships among the items on a scale. Methods of assessing unidimensionality using factor analysis include the chi-square test, which must be used with larger samples (Bock & Lierberman, 1970; Joreskog, 1978).

Other factor analysis-based methods of establishing unidimensionality involve communalities, which are the proportion of an item's variance that be explained by the underlying trait or component. Communality indices examine the loadings of items or variables on a component in relation to the square root of the items' communalities, as well as the correlation between items (Green, Lissitz, & Mulaik, 1977; Hattie, 1985).

There exists a multitude of ways to establish dimensionality or the number of factors, and "the simultaneous use of multiple decision rules is appropriate and often desirable" (Thompson & Daniel, 1996, p. 200). Given the variety of options of assessing

unidimensionality and the evaluation of the methods throughout the literature, the eigenvalue ratio rule established by Lord (1980) was used to argue essential unidimensionality of the SAS. Results from the dimensionality analysis were used to establish evidence for *Standard 1.13* (AERA, APA, & NCME, 2014).

Measurement Theory

Researchers have used several analysis as effective methods for establishing the reliability and validity of a construct, including classical test theory (CTT), item response theory (IRT), and structural equation modeling (SEM), which includes latent change modeling and confirmatory factor analysis (CFA). Though CTT has been used in educational and psychological measurement for nearly a century, IRT and SEM – including its subsumed models – have experienced a growth in use over the past several decades. The CTT measurement framework is flexible in that CTT requires few theoretical assumptions which are difficult to violate, and thus is applied to a wide variety of assessments. CTT primarily focuses on providing information at the test level in reliability, though CTT also provides item-level data, such as difficulty and discrimination. However, statistics derived from CTT are sample dependent. New measurement models such as IRT have improved upon some of the weaknesses inherent in CTT, though the two approaches are complementary.

Item response theory (IRT) was initially developed for educational assessment. IRT does not assume that reliability or errors are evenly distributed, and it places both items and people on the same scale (theta, or the latent trait being measured). In addition to differentiating reliability across a scale, IRT also acknowledges that some items more

accurately measure a trait at certain points on the scale (Hambleton & Swaminathan, 1985). This advantage allows those using IRT to construct shorter scales than is possible with CTT (Hambleton & Jones, 1993). Though IRT was developed in educational testing, IRT has been used in fields like psychology (Winterstein, Ackerman, Silvia, & Kwapil, 2011; Embretson, 1996). The next sections discuss the major assumptions and concepts unique to IRT, as well as the three main IRT models addressed in this paper.

Local Independence

One of the main assumptions of IRT that is not present in CTT is the assumption of local independence, which is closely related to the assumption of unidimensionality. Hambleton and Swaminathan (1985) described local independence as occurring when "an examinee's responses to different items in a test are statistically independent" (p. 23). Specifically, a person's response on one item has no effect on their response to any other item when the parameters are held constant, meaning that any two items on a test or scale are uncorrelated after accounting for the latent trait (θ). The probability of a person answering two items correctly, or endorsing two items, are independent of one another after one removes the effect of θ. To understand the probability of the person answering both items correctly, one must take the product of the two probabilities of answering each item correctly. If, however, the probability of one item correctly is contingent upon answering another item correctly, the items are not locally independent. Local independence can be represented as

$$P(U_i \ldots U_n|\theta) = \Pi_i P_i(\theta)^{u_i} Q_i(\theta)^{1-u_i}$$

where *i* denotes items, *P* is the probability of correctly answering a certain item given someone's θ, *Q* is the probability of incorrectly answering an item, and *U* is the response pattern for a set of items (Hambleton & Swaminathan, 1985).

Local independence applies to people as well. Given two examinees' θ values, their probabilities of answering an item a certain should be uncorrelated (Allen & Yen, 1979, p. 241). Like the local independence of items, the local independence of examinees states that the probability of two examinees answering the same item correctly is the product of their individual probabilities.

Local independence is directly related to the assumption of unidimensionality (Hambleton & Swaminathan, 1985). For example, because the SAS purports to measure the latent trait of ambivalence, all items should measure ambivalence to some degree. Thus, the items are correlated when considering the construct being measured by the SAS. However, after removing the construct being measured, the items should not be correlated with one another. If the items are correlated after removing the ambivalence construct, then the items violate the assumption of local independence.

Violations of local independence can occur for many reasons, such as practice effects or fatigue, as Yen (1993) identified. To violate the assumption of local independence, consider an example using the SAS. If one question on the SAS cued a person to answer the following question a certain way after removing the effect of ambivalence, the latent trait, the assumption of local independence would be violated. Specifically, this means that person's answer on one question was dependent on his or her answer on another question.

However, the assumption of local independence can hold when data are not unidimensional. For example, a multidimensional scale assessing three latent traits could still meet the assumption of local independence if the items are independent after accounting for the effects of all three traits. With this multidimensional data, items can measure multiple latent traits, or each item can measure one trait. Regardless, local independence under multidimensional data can only be evaluated after accounting for the effects of all latent traits measured.

The assumption of local independence can easily be checked in IRT by simple pairwise comparisons using the χ^2 statistic. This includes Lord's (1953) χ^2, which was later expanded upon by Chen and Thissen (1997) through their Pearson χ^2. Lord's (1953) approach examines the significance of the χ^2 resulting from multiple χ^2 of pairs of items on scale. The χ^2 statistic is calculated for each pair of items across ability levels. Significant χ^2 statistics indicate items that researchers should investigate for violation of the assumption of local independence.

Another approach to testing the assumption of local independence is Yen's (1993) Q_3. Yen's approach examines the correlation of each pair of items on a test after controlling for performance on all other items (Mislevy, Rupp, & Harring, 2012; Yen, 1993). A Q_3 of around 0 is indicative that the assumption of local independence holding, as was as the assumption of unidimensionality. A rule of +/- 0.20 is used as the cut off for assuming local independence (Yen, 1993). A Q_3 with a difference from 0 greater than 0.20 is indicative of an issue with local independence and dimensionality.

Another method of assessing local independence of items is the Jackknife Slope Index (JSI; Edwards, Houts, & Cai, 2017). FlexMIRT software (Cai, 2013) outputs crosstab matrices of the JSI which is based on the notion that item pairs that are not locally independent have steep slopes. Edwards, Houts, and Cai (2017) reported seeing such patterns in their research. Thus, the JSI index is intended to flag item pairs with steep slopes, and a high JSI indicate item pairs that may violate the local independence assumption. While there is not a currently established metric or threshold for what constitutes as a large JSI index, indices that appear to be outliers generally warrant investigation (Edwards, Houts, & Cai, 2017).

One last method of establishing local independence, and which also establishes a case for essential unidimensionality, was created by Stout (1987). He specifically set out to create a statistical test, DIMTEST, that would assess the essential unidimensionality of a scale with dichotomous items. DIMTEST examines unidimensionality by first establishing local item independence, the notion that a person's responses on one item are not impacted by their responses on another item after accounting for the latent trait (Nandakumar, 1993; Nandakumar & Stout, 1993). DIMTEST splits the sample into separate groups and examines the independence of the samples in those groups. Specifically, DIMTEST uses an EFA to automatically separate items into an assessment subtest (AT), where all items in the AT group measure a similar factor. It is also possible to do separate items into AT by a confirmatory method, in which the user selects which items to group in AT. Then, all items not placed in the AT group are placed in the partitioning subtest (PT) group, which is based on participants' scores (Stout, 1987).

Participants are broken out into an unnamed number of groups per their scores on items in the PT group (Nandakumar & Stout, 1993). Then, the average participant score by group is computed, as well as the variance in scores by group. DIMTEST provides a *T*-statistic that compares the variation of scores by group, and this *T*-statistic is interpreted using a *p*-value. A *p*-value that is not significant indicates that the group variances are similar, and thus, essential unidimensionality holds. DIMTEST has been widely used (Meara, Robin, & Sireci, 2000; Nandakumar & Feng, 1994) and established as a reliable method of both establishing local independence and detecting unidimensionality (Hattie, Krakowski, Rogers, & Swaminathan, 1996).

DIMTEST is proven to perform two purposes: assessing the assumption of local independence and simultaneously assessing the assumption of essential unidimensionality. The DIMTEST procedure is also based in IRT, and one of the purposes of this paper is to provide an example of novel circumstances in which IRT approaches can be applied. Thus, DIMTEST was used to assess local independence, and to confirm essential dimensionality tested through the eigenvalue ratio method (Lord, 1980). The JSI index was also used to strictly assess local independence.

Information

In CTT, reliability is defined as the internal consistency of a measure; or the extent to which items perform similarly across various instances. Reliability can also refer to indices uses to measure internal consistency, such as Cronbach's α. Reliability is especially important in high-stakes testing and in instances where psychological diagnoses are made. Ideally, assessments that result in diagnoses exhibit high reliability.

A weakness of reliability in the context of CTT, however, is that it assumes that reliability is the same across items and scores on any scale.

IRT improves upon the reliability concept with information, which acknowledges that information may change across the scale as a function of θ levels and allows for more precision at different points on the θ. The next two sections discuss information at the item level and at the test level.

Item information. Unlike CTT, IRT allows one to estimate the contribution of each item to the overall test, and this amount of contribution varies across items. Each item provides an item information function (IIF), which indicates the amount of information provided by an individual item. The information provided by an individual item is based on "the slope of the item response function" (Hambleton & Swaminathan, 1985, p. 105), which is a function of the item parameters estimated through a model. If more information is needed at a certain point in the scale, such as at a cut point, one may add items around that point. Similarly, items close to a person's θ provide more information that items that are not close to a person's θ. Ultimately, the IIF at point on the θ scale can be added to arrive at the test information function.

Test information. The test information function (TIF) is a result of the contributions of each item. Test information allows one to consider the amount of information at each point on the theta scale (θ) to determine precision at each ability θ. Test information is impacted by the number of items on a scale, by the quality of items on a scale, and by difficulty and discrimination of those items in relation to the θ levels of

the respondents. The TIF varies with θ, and as information increases near a θ, then so does the height of the test information function (Hambleton & Swaminathan, 1985).

Measurement error is inversely related to information, so the least amount of error occurs where the most amount of information occurs. Standard error is indicated as

$$SE(\theta) = \frac{1}{\sqrt{I(\theta)}}$$

where the standard error is the lower asymptote of the information function. Thus, the standard error has an inverse relationship with information at a certain θ level. This means that the most precision occurs where the most amount of information is located, and that one can have confidence in the scores around that θ. Results from evaluating item information and test-level information will provide "appropriate evidence of reliability/precision" (*Standard 2.0*; AERA, APA, & NCME, 2014, p. 42), evidence for the reliability of scores (*Standard 2.3*), and evidence for use of scores in making predictions (*Standard 7.12*).

The notion that test information might be higher at a certain point on the θ scale than at other points in the scale is rooted in the concept of precision. High-stakes assessments and psychological scales used to make clinical decisions should have the most precision, or information, at the cut score point. Test information, in conjunction with item information, allows one to target a point on the θ scale where the most precision is needed. Thus, items that are more precise around a cut point might be added to a test, and items that provide minimal contributions to information at that cut point

might be discarded. However, item information is dependent on the item parameters estimated, such that item and test information functions vary depending on the model used.

Models

One can include up to three parameters using logistic IRT models, using three different yet hierarchical models: 1 parameter-logistic (1PL)/Rasch, 2 parameter-logistic (2PL), and 3 parameter-logistic (3PL). The 1PL model includes the difficulty parameter (*b*). The 2PL model, developed by Birnbaum (1968) and which has seen more use in psychological measurement than more complex models (Edwards, Houts, & Cai, 2017), includes the difficulty parameter and adds a discrimination parameter (*a*); the discrimination parameter is comparable to point biserial statistics in CTT. This discrimination parameter in the 2PL model indicates the degree to which an item differentiates across people at various levels of a trait. The higher the a-parameter (0-3), the more an item discriminates across people possessing different levels of a trait. Finally, the 3PL model includes the guessing parameter (*c*) in addition to parameters *a* and *b*, which takes into the account the possibility of someone falsely endorsing an item; the *c*-parameter can also consider the degree to which someone low on a trait falsely endorses an item with a higher b-parameter.

The 3PL model is written as such:

$$P_i(\theta|a,b,c) = c_i + (1 - c_i)\frac{e^{a_i(\theta - b_i)}}{1 + e^{a_i(\theta - b_i)}}$$

where *i* denotes the item (Hambleton & Swaminathan, 1985). If the *c* parameter is constrained to 0 in the above model, it becomes the 2PL model. If the *b* parameters are also constrained to be equal, it becomes the 1PL model.

All models generate estimated item characteristics curves (ICCs). ICCs display the probability of someone endorsing an item across an ability level. The ICCs are derived from the IRT model chosen, so the ICCs are sensitive to each IRT model's parameters. For example, ICCs in the 1PL model are affected only by the *b* parameter, so the ICCs for each item vary only by ability level. However, item data model via the 2PL model are affected by both the *b*- and *a*- parameter. Items with a higher *a* parameter, or those items that are more discriminating, yield steeper slopes on the ICCs, so the ICCs vary with ability and discrimination. Under the 3PL model, items are affected by all three parameters, where the *c* parameter is distinguished through the lower asymptote. Because guessing is evaluated under the 3PL model, the probability of not endorsing an item, or answering an item incorrectly, is no longer zero.

As alluded to earlier, all three models have an impact on information, as the amount of information depends on the model used and the estimated ICCs derived from the model. Thus, as the model and number of parameters change, so does information. For 1PL models, "the maximum value of the information is constant" across items

(Hambleton & Swaminathan, 1985, p. 105). However, for 2PL models, information increases as the discrimination parameter increases. Highly discriminating items should naturally provide more information. Lastly, information increases as the guessing parameter, c, decreases under 3PL models. This result also makes sense, as a high *c*-parameter indicates a higher probability of someone falsely endorsing an item or answering an item correctly when they should not. As one might expect, an item with a high c parameter would provide very little information about one's ability level.

Model fit and utility. The use of an IRT model depends on the data and model fit. Model fit is assessed by comparing the data to expected values for the model (Maydeu-Olivares, 2013). Several goodness of fit statistics exist that can be applied to assess the discrepancy between the model and the data, and there is not universal agreement on which are best, so researchers usually report more than one. Chi-square statistics are often used, but χ^2 is sensitive to large data sets, meaning that any minor difference between the data and model results in a significant χ^2. Other goodness-of-fit statistics that are more robust than χ^2 include the root mean square error of approximation (RMSEA; Maydeu-Olivares, 2013), the Akaike information criterion (AIC; Akaike, 1974), and the Bayesian information criterion (BIC; Schwarz, 1978).

Fit can also be assessed at the item level. One method of assessing fit is conducting a residual analysis. A residual analysis compares the rate of endorsement for an item at a certain ability level and compares it with the predicted rate of endorsement for that item. These are usually analyzed visually with estimated item characteristics curves (ICCs) and corresponding residual plots. Chi-square item fit statistics are also a

popular method used to interpret item fit (Orlando & Thissen, 2000). There are several methods, however, Orlando and Thissen's (2000) fit statistic is preferable as it "is based on test scores" rather than "model-dependent ability estimates" (Kang & Chen, 2008, p. 393).

In choosing a model, Hambleton and Swaminathan (1985) suggest that "fitting more than one model and comparing (for example) the residuals provides information that is invaluable in determining the usefulness of item response models" (p. 168). Hambleton and Swaminathan (1985) further suggest that multiple methods of analysis and the evaluation of models is related the importance of the research being conducted (p. 168). It is clear that no one fit statistic is best; rather, it is the confluence of different analyses that indicate both the fit of the model to the data, and more importantly the utility of the model to the data. A goodness-of-fit statistic might indicate that the model fits the data well, but it is the additional analyses of the item statistics, derived from both CTT and IRT, that provide information on the utility of the model for the data.

Both RMSEA and AIC goodness of fit statistics were used to compare the three models. After a model was chosen, model utility was evaluated via item parameters as generated through CTT and IRT. Cronbach's α, the TIF, and IIFs were evaluated to assess reliability and information provided by the SAS. Given that the SAS is used to measure one's level of ambivalence, a trait identified in those at-risk for schizophrenia, it is paramount that the scale is not only reliable, but that the most confidence in scores occurs at the cut point.

Results from IRT and CTT fit and utility analyses provided evidence for *Standards* related to reliability, validity, and documentation (AERA, APA, & NCME, 2014). Specifically, results from both the model fit analysis and analysis of items provide evidence for the internal structure of the data (*Standards 1.13* and *2.5*), evidence for performance on the SAS at a given level (*Standard 1.18*), and evidence that the psychometric properties of the SAS have been evaluated (*Standard 4.10*).

Differential Item Functioning

One of the assumptions of IRT is that the item parameters are invariant across different groups. Differential item functioning (DIF) is one method used to investigate invariance, and is a term used to describe a type of test unfairness. DIF is important to investigate because items exhibiting DIF have construct-irrelevant characteristics is interfering with individual performance. When two subgroups with the same level of an underlying trait (θ) have a different probability of endorsing an item, DIF is occurring (Smith & Reise, 1998). However, one must differentiate DIF from impact, as Dorans and Holland (1993) state:

> Impact refers to a difference in performance between two intact groups. Impact is everywhere in test and item data because individuals differ with respect to the developed abilities measured by items and tests, and intact groups, such as those defined by ethnicity and gender, differ with respect to the distributions of developed ability among their members. (p. 36)

Thus, there are instances in which subgroups that have differing levels of a trait perform differently on an item, but those items do not exhibit DIF. For example, reports of somatic depressive symptoms occur at a higher rate in females than in males

(Silverstein, 2002), so the two groups possess different levels of the somatic depression trait. If, after controlling for the underlying trait of somatic depression, females and males endorse an item on a scale assessing somatic depression at a similar rate, DIF is not occurring. However, if after controlling for somatic depression, the two groups endorse that item at different rates, the item is exhibiting DIF. Essentially, DIF can occur if two subgroups perform differently on an item after accounting for the latent trait(s) of interest. If one found an item or items to be exhibiting DIF, this construct-irrelevant interference would be a threat to valid interpretation of scores across these groups. Indeed, the *Standards* dedicate a chapter to fairness in which *Standards 3.0* and *3.1* reference the importance of removing construct-irrelevant variance and ensuring valid interpretations for all subgroups of a population (AERA, APA, & NCME, 2014).

Most often, gender and ethnicity are investigated as sources of DIF, but other factors, such as comorbid psychological disorders, can be sources of DIF as well. One should examine the presence of DIF across all items to ensure that construct-irrelevant factors are not influencing the results of answers to a scale or test. If there are factors outside of the construct being measured affecting one's score, this negatively impacts the validity of the score, as well as the interpretability of one's score. If a score on the SAS is no longer indicative of how much ambivalence one possesses, to use a score on the SAS as a criterion by which one begins treatment would be a misinterpretation of the score. DIF, naturally, is related to both dimensionality and local independence.

Through DIF, one group is identified as the reference group, and the other group is identified as the focal group; the responses of these two groups on the same items are

then compared to see if people of the same ability level from the two different groups are performing differently on certain items. The reference group is the group to which the focal group is compared; the focal group is usually the group believed to be disadvantaged. DIF can be analyzed through two types of methods: observed score methods and latent trait methods (Apinyapibal, Lawthong, & Kanjanawasee, 2015), the former of which is the focus in this study. Observed score methods include the Mantel-Haenszel statistic (MH; Hambleton, Swaminathan, & Rogers, 1991). Under this method, the scores of two subgroups are compared using contingency tables, resulting in a statistic, α_i, which is a ratio of the probability that a reference group has endorsed an item more frequently and probability that a focal group has endorsed an item more frequently. However, this α_i is usually interpreted through a delta effect size statistic, known as MH D-DIF. Depending on the magnitude, the absolute value of the MH D-DIF index is placed into one of three categories: A (negligible DIF), B (moderate DIF), and C (large DIF; Zieky, 1993). Items in category A do not have a significant value, or exhibit negligible DIF; items in category B are significantly different from zero, or exhibit moderate DIF; and items in category C are significantly different from zero and are greater than 1.0, or exhibit large DIF. Negative MH D-DIF effect sizes indicate that the reference group is favored, and positive MH D-DIF effect sizes indicate that the focal group is favored.

Another well-known index is Simultaneous Item Bias (SIBTEST; Shealy & Stout, 1993), which functions similarly to MH DIF. In addition to producing a significance test, a beta-uni (β_{uni}) index is produced as well, which functions as an effect size. This effect

size indicates the magnitude of DIF. Stout and Roussos (1996) place the indices into one of three categories: A, in which the absolute value of the β_{uni} index is not greater than .059; B, in which the absolute value of the β_{uni} index is between .059 and .088; and C, in which the absolute value of the β_{uni} index is higher than .088 (Stout & Roussos, 1996). Negative values indicate DIF in favor of the focal group; positive value indicate DIF in favor of the reference group. Like MH DIF, the categories are defined as negligible, moderate, and significant. Results from the MH DIF and SIBTEST methods are typically comparable, though SIBTEST can detect DIF better than MH DIF under certain conditions (Stout & Roussos, 1996).

Assessment of DIF can add credibility to the interpretation of results from a psychological scale. For example, Emmert-Aronson, Moore, and Brown (2014) conducted a DIF analysis of items on the Anxiety Disorders Interview Schedule for DSM-IV: Lifetime (ADIS-IV-L, Di Nardo, Brown, & Barlow, 1994), which assesses major depressive disorder (MDD), among other mood disorders. Researchers found that none of the symptoms of MDD assesses through items on the ADIS-IV-L exhibited DIF, confirming that that use of those items measuring MDD is consistent across subgroups.

Detection of DIF is also useful in flagging potentially problematic items. Waller, Compas, Hollon, and Beckjord (2005) investigated DIF on the Beck Depression Inventory-11 (BDI-II) for women with breast cancer and women with clinical depression. Fifteen of the 21 items on the BDI-II exhibited DIF, indicating that endorsements from either group were not indicative of the underlying depression trait. Similarly, a cross-cultural study of the BDI-II (Canel-Cinarbus, Cui, & Lauridsen (2011) found that 12

items exhibited DIF. Though Canel-Cinarbus et al. (2011) suggest that some of the DIF could be attributed to the translation of the BDI-II from English to Turkish, 11 of the 12 items identified as exhibiting DIF were also found to exhibit DIF in the Waller et al. (2005) study.

Given the importance of DIF in establishing the validity, use, and interpretability of scores, the presence of DIF on the SAS was evaluated in terms of sex. Both the MH procedure and the SIBTEST procedure were used to investigate DIF across items on the SAS. These analyses provided evidence for *Standards 3.0* and *3.1* (AERA, APA, & NCME, 2014).

Summary

Given that high scores on the SAS are used as a predictor of later development of schizotypy and schizophrenia, precision of the SAS and interpretability of the scores that result impact those to whom the SAS has been administered. Further, interpretations of these scores is essential to accurately understanding one's predisposition for later development of these disorders. Extensive use was made of the relevant analyses via IRT and CTT to provide a detailed evaluation of the overall validity and reliability of the SAS, as well as the interpretability of scores that result from the SAS. Such an evaluation provides those in psychological measurement and those who study and use the SAS with information regarding the internal structure of the SAS, the characteristics and quality of the items, and the point at which the SAS provides the most information. Further, tying the results of the analyses to the *Standards* provides the SAS with stronger documentation and evidence accepted by experts in the field.

CHAPTER III

METHODOLOGY

Few studies have been conducted to analyze the quality of the SAS, and few studies have been conducted on psychological scales using IRT and its range of analyses. Given the importance of the role of the ambivalence construct to the concepts of schizotypy and schizophrenia, and the need for more research on the topic of ambivalence and the SAS as indicated by gaps in the literature, a thorough evaluation of the SAS through a dimensionality analysis, CTT, IRT, and DIF were conducted to illustrate the uses of these methods in a field other than education.

This chapter is organized first by describing the sample used for this study, followed by descriptions of the analyses aligned to each research question. Under each research question, the analyses introduced earlier are further described, along with the applicable software programs used to conduct each analysis. In addition, each section describes evidence of the relevant *Standards* (AERA, APA, & NCME, 2014) that each analysis supports.

Sample

This study used existing data to evaluate the properties of the SAS. The psychology department at the University of North Carolina at Greensboro (UNCG) has a database of undergraduate students who have been administered multiple psychological scales over nearly 10 years, one of which includes the SAS. The number of participants

who have been administered the SAS exceeds 7,000 subjects, which is more than sufficient for IRT, CTT, and PCA analyses. Participant responses to the SAS were used to answer the research questions. The use of the existing SAS database was identified as exempt by the UNCG Institutional Review Board and approved for use in this study (see Appendix B).

While the dataset was not representative of a clinical sample, the dataset was appropriate for use in analyzing the properties of the SAS. The onset of schizophrenic disorders is usually early adulthood – and the SAS was administered to a large sample of college students who typically fall within that age range. Thus, the university sample was both large enough and appropriate for the purpose evaluating the properties of the SAS.

Research Questions

To thoroughly evaluate the properties of the SAS, the following questions were addressed to establish the best fitting model:

1) What is the dimensionality of the SAS?
2) Which IRT model best fits the data?

Once the dimensionality of the SAS was determined and the appropriate IRT model was chosen, the following research questions related to the psychometric properties of the SAS were addressed:

3) Do the item characteristics meaningfully explain the data?
4) Which SAS items provide the most information? Does the most information occur near high scores on the SAS?
5) Do any items exhibit DIF?

Table 1 displays the research questions, the methods and analyses by which each research question was addressed, and the evidence for the relevant *Standards* (AERA, APA, & NCME, 2014) each analysis supported.

Table 1. Analysis and Evidence Matrix

Research Question	Analyses	Analysis Approach	*Standards* Evidence
1. What is the dimensionality of the SAS?	Establish essential unidimensionality	a. Ratio of the first and second eigenvalues is 4:1 (Lord, 1980) or 3:1 (Slocum-Gori & Zumbo, 2011) b. DIMTEST (Stout, 1987) to confirm results from eigenvalues comparison	*Standard 1.13*
2. What IRT model best fits the data?	Compare the fit of 1PL, 2PL, and 3PL models	a. Comparison of the RMSEA and AIC statistics reported out from FlexMIRT for all models	*Standard 1.13, Standard 4.10*
Research questions following model selection			
3. Do the item characteristics meaningfully explain the data?	Compare item fit and statistics for the chosen model	b. Evaluate item difficulty through CTT *p*-values and IRT *b*-parameters c. Evaluate item discrimination through CTT point biserials and IRT *a*-parameters, if applicable d. Evaluate item fit through item level χ^2 statistics e. Evaluate local independence through JSI and DIMTEST	*Standard 1.13, Standard 4.10*

Research Question	Analyses	Analysis Approach	*Standards* Evidence
4. Which SAS items provide the most information? Does the most information occur near high scores on the SAS?	Evaluate item and test-level information	a. Generate IIF for all items b. Generate TIF c. Calculate Cronbach's α	*Standard 2.0, Standard 2.3, Standard 7.12*
5. Do any items exhibit DIF?	Evaluate DIF for all items	a. SIBTEST and MH DIF analyses	*Standards 3.0, 3.1*, and *4.10*

What is the Dimensionality of the SAS?

As described in the previous chapter, there exists a multitude of ways to evaluate the dimensionality of a scale or test, and some are better than others. To establish essential unidimensionality, the relationship of eigenvalues is often used; one of the more accepted methods of using eigenvalues to establish essential unidimensionality is the ratio of the first and second eigenvalues, specifically where the ratio is 4:1 (Lord, 1980) or the ratio is 3:1 (Slocum-Gori & Zumbo, 2011). Thus, the ratio of the first and second eigenvalues was evaluated to establish essential unidimensionality. If the eigenvalue ratio approach 4:1, essential unidimensionality could be claimed.

However, given Hattie's (1985) critique of solely using eigenvalues to establish essential unidimensionality, DIMTEST was used to confirm or refute results from the eigenvalue ratio method. Stout's (1987) method evaluates dimensionality based on local independence, so it also served to confirm results from the local independence statistics produced to answer research question 3. To conduct the analysis, DIMTEST Version 2.1 (Stout, 2006) was used. To establish the AT group, the sample was first split in half. An EFA was conducted on one half of the sample to determine the AT items; the second half of the sample was used to run the DIMTEST analysis by specifying the AT items determined through the EFA. This approach reflects best practice to avoid inflated Type 1 error rates (DIMPACK Version 1.0; Stout, 2006).

If, the results from the eigenvalue ratios and DIMTEST provided conflicting evidence, the eigenvalue ratios were used, but local independence analyses were

evaluated to uncover whether any issues in local independence were affecting the dimensionality of the SAS.

Results from this analysis aligned to *Standard 1.13* (AERA, APA, & NCME, 2014), which suggests that "if the rationale for a test score interpretation for a given use depends on premises about the relationships among test items or among parts of the test, evidence concerning the internal structure of the test should be provided" (pp. 26-27). Moreover, to claim that the SAS exhibits essential unidimensionality and thus, only measures the ambivalence construct, support for such a claim must be provided.

What IRT Model Best Fits the Data?

Once essential unidimensionality was established, the SAS data were modeled via 1PL, 2PL, and 3PL models. These models are reasonable to use, and the 2PL model has "seen widespread use in psychological measurement" (Edwards, Houts, & Cai, 2017, p. 2). The data was modeled using FlexMIRT software (Cai, 2013). FlexMIRT software is a syntax program that can be used to fit multi-dimensional and unidimensional IRT models, to evaluate item parameters, to evaluate the assumption of local independence, and to evaluate item and model fit.

To assess the goodness of fit of the 1PL, 2PL, and 3PL models, the RMSEA and AIC goodness of fit indices were calculated to assess the fit of the model used. Upon modeling the data via the 1PL, 2PL, and 3PL models, the researcher compared the IRT models with the RMSEA index, in which a value of .07 (Steiger, 2007) or below indicates good fit. The AIC fit index was also interpreted, where a lower AIC value indicates better fit. The fit statistics were reported by model and statistic.

Analyses from this research question further contribute evidence to the structure of the SAS, or *Standard 1.13*. Furthermore, the analysis of which model best fits the data provides evidence in support of *Standard 4.10*, which requires that the model used to evaluate the psychometric properties of items and an assessment be documented (AERA, APA, & NCME, 2014, p. 88).

Do the Item Characteristics Meaningfully Explain the Data?

The degree to which item characteristics meaningfully explain the data can be assessed in a variety of ways. The utility of a model is assessed based on item parameters, item fit statistics, evaluation of local independence, and whether the statistics make sense given the construct and purpose of the scale. Once a model is chosen, the item parameters were examined to assess model utility, or how well the item statistics explain the SAS data. Using SPSS Version 19 (IBM Corp, 2010), CTT was used to generate *p*-values and point-biserial correlations, and FlexMIRT (Cai, 2013) was used to generate relevant IRT item-parameter statistics. CTT and IRT item characteristics were summarized together to highlight any similarities or differences between the two methods. The ICCs were generated to evaluate the relationship between each item and the underlying construct. Additionally, item-level χ^2 fit statistics (Orlando & Thissen, 2000) were calculated in FlexMIRT (Cai, 2013) to evaluate item-level fit.

Finally, the JSI index was used (JSI; Edwards, Houts, & Cai, 2017) to evaluate local independence. The JSI index is a newer index based on the idea that item pairs that are not locally independent have steeper slopes than items pairs that are locally independent. Larger JSI indices indicate item pairs that may violate the local

independence assumption; thus, larger JSI indices were flagged as item pairs possibly violating local independence. As there currently is not a threshold or rule of thumb for large JSI index values, larger JSI indices were identified by comparing JSI indices across the pairs of items and noting any outliers. FlexMIRT (Cai, 2013) was used to generate the JSI indices matrix.

The analyses to identify whether the item characteristics meaningfully explain the SAS data provide further evidence in support of *Standard 1.13* (AERA, APA, & NCME, 2014) which requires evidence of the internal structure of the test, and *Standard 4.10,* which requires that the psychometric properties such as IRT model, item parameters, and item characteristics be documented.

Which SAS Items Provide the Most Information? Does the Most Information Occur Near High Scores on the SAS?

To evaluate the point at which the most information occurs for the SAS, Cronbach's α was calculated via SPSS version 19, and the IIFs and TIF through R 3.4.0 (R, 2017) using the output files from FlexMIRT (Cai, 2013). To evaluate item-level information, the IIFs were used to evaluate which items provide the most information at the higher end of the θ scale. The IIFs revealed which items, if any, provided little information at the higher end of the θ scale. The researcher used the TIF to identify the point of maximum information on the θ scale. Test-level reliability and information were summarized through Cronbach's α and the TIF and corresponding SEM.

The *Standards* (AERA, APA, & NCME, 2014) suggest that one must provide evidence of reliability or precision of an instrument, and that one should document the

methods used to establish reliability or precision. Producing Cronbach's α, as well as IIFs and the TIF provide evidence in support of *Standard 2.0,* which requires evidence of reliability for the interpretation of scores; *Standard 2.3,* which requires that reliability procedures align to the internal structure of the test; and *Standard 7.12,* which requires evidence for use of scores in making predictions.

Do any Items Exhibit DIF?

For the last analysis, DIF was evaluated in relation to sex. For this analysis, male participants were treated as the reference group and female participants were treated as the focal group. Despite results from Mann et al. (2008), DIF was investigated for sex, as the researcher was curious whether sex differences in the prevalence of schizophrenia would translate to sex differences in the performance on the SAS. DIF was identified through the use and comparison of the MH D-DIF indices and SIBTEST (Shealy & Stout, 1993). The MH D-DIF statistic was calculated through the R program difR (Magis, Beland, Tuerlinckx, & De Boeck, 2010), and SIBTEST DIF statistics were produced through the R program mirt (Chalmers, 2012). These statistics were used to flag any potentially problematic items and compare the classification of the items across the three levels of DIF – negligible, moderate, and large/significant. Potentially problematic items were those items that fall into the moderate or large/significant categories.

Results from these analyses provided evidence for the psychometric properties and the fairness of the SAS. Specifically, the analysis of the presence of DIF provides additional evidence for *Standard 4.10* (AERA, APA, & NCME, 2014) which requires that statistics such as DIF be documented. Results also provided evidence for *Standard*

3.0, which details the importance of detailing the processes used to remove construct-irrelevant characteristics, as well as *Standard 3.1*, which requires that "test developers need to be knowledgeable about group differences that may interfere with the precision of scores and the validity of test score inferences" (AERA, APA, & NCME, 2014, pp. 63-64).

CHAPTER IV

RESULTS

This section presents the results from the CTT and IRT analyses described in Exhibit 1. Results are organized by research question. Under each research question follows a brief reference to the methodology and software program, if applicable, along with a description of the results.

What is the Dimensionality of the SAS?

Dimensionality of the SAS was evaluated in SPSS Version 19 via a PCA, which provides eigenvalues for each factor or dimension. Results from the PCA analysis yielded an initial eigenvalue of 4.98, and a second eigenvalue of 1.32. Table 2 displays the components extracted from the PCA and the corresponding eigenvalues and variance explained.

Table 2. SAS PCA Eigenvalues

Component	Eigenvalue	% of Variance
1	4.98	26.22
2	1.32	6.94

The ratio of the first to second eigenvalue is approximately 3.78:1, which falls in between the 4:1 benchmark (Lord, 1980) and the 3:1 benchmark (Slocum-Gori & Zumbo,

2011). Use of the eigenvalue ratio rule indicated that the data can be interpreted as exhibiting essential unidimensionality.

However, to further confirm whether the SAS is essential unidimensional, DIMTEST was also used. DIMTEST results conflict with results from the analysis of eigenvalue ratios. After conducting an EFA on half of the SAS sample, the software determined that items 12, 13, 14, and 16 comprised the AT group. This AT group was used in the other half of the SAS sample in conducting the DIMTEST analysis. The results of that analysis indicate that the data may not be unidimensional, $T = 7.44$, $p < 0.01$.

Given the contradictory results, a PCA on the tetrachoric correlation matrix of the SAS data was conducted via the program psych (Revelle, 2017), as PCAs using tetrachoric correlations are appropriate for use with dichotomous data. Within the program psych is principal, which returns only the best components. The program rescales the eigenvectors to return loadings similar to a factor analysis. Results indicated that one component was sufficient to explain the SAS data, and all component loadings were greater than 0.55, providing evidence in support of essential unidimensionality.

Results from the eigenvalue analysis suggest that the data can be interpreted as essential unidimensional; however, results from the DIMTEST analysis suggest that there may be multiple underlying factors or issues with local independence. Given the adequate results of the eigenvalue analysis, the results from the PCA on the tetrachoric correlation matrix, and the prevalence of Type 1 error rates when using DIMTEST, a unidimensional

IRT model was chosen to model the data. However, DIMTEST results were also used in evaluating the SAS item pairs for violations of local independence.

What IRT Model Best Fits the Data?

Once the SAS was determined to be essential unidimensional via the eigenvalue ratios of 4:1 and 3:1, the data could be represented through a unidimensional IRT model. FlexMIRT (Cai, 2013) was used to calibrate the 1PL, 2PL, and 3PL models to the data and to produce the RMSEA and AIC model fit statistics. FlexMIRT software calibrates a 2PL model by constraining a graded response model (Samejima, 1969) to have two categories, which is equivalent to a 2PL model (Cai, 2013; Houts & Cai, 2013).

Table 3 displays the fit statistics for all three models.

Table 3. Fit Statistics for the 1PL, 2PL, and 3PL Models

Fit Statistics	1PL	2PL	3PL
RMSEA	0.03	0.03	0.03
AIC	142250.66	141997.68	142158.58

Across models, the RMSEA remained constant with a value of RMSEA = 0.03, indicating good fit. Further examination of the AIC revealed that the 2PL best fit the data, as the AIC was lowest for this model (AIC = 141997.68). Thus, the 2PL model best fit the data and was used to evaluate item parameters, ICCs, and item fit statistics.

Do the Item Characteristics Meaningfully Explain the Data?

As the 2PL model best fit the data, the item parameters included both difficulty, or rate of endorsement, and discrimination. Thus, the location of each item on the θ scale was determined by two item parameters, *b* and *a*. Establishing whether the item

characteristics meaningfully explain the data is multi-faceted. This process involves evaluating item parameters such as difficulty or rate of endorsement, discrimination, but also local independence and item fit statistics. To analyze item-level characteristics, the researcher first computed CTT statistics in SPSS Version 19 and the 2PL item parameters. Table 4 displays item-level statistics.

Item-level statistics show that participants endorsed items 4, 8, 10, and 15 at a higher rate than other items, which had *b*-values of -0.78, -0.07, -0.42, and -0.33, respectively. Participants endorsed items 1, 6, 11, and 14 at the lowest rates, which had *b*-values of 1.45, 1.51, 1.45, and 1.37, respectively. The discrimination parameters show that items 2, 3, 9, 14, and 19 discriminate the most across the θ scale, which had *a*-values of 1.64, 1.57, 1.58, 1.54, and 1.56, respectively; in most cases, the point-biserial values for these items confirmed these conclusions. In general, point-biserial values of approximately 0.30 or higher are reasonably good; all items exceeded this threshold.

As a note, FlexMIRT (Cai, 2013; Edwards & Cai, 2013) does not include the *D* constant in the models; rather, this value (-1.7) is absorbed by parameters, which could make them lower than they appear. Still, *a*-parameter values of 1.0 or greater typically indicate adequate discrimination; thus, the *a*-parameters are reasonable.

Table 4. Item Difficulty and Discrimination

	2PL		CTT	
Item	*b*-value	*a*-value	*p*-value	Point biserial
1	1.45	1.44	0.18	0.41
2	0.84	1.64	0.28	0.49
3	0.45	1.57	0.38	0.49
4	-0.78	1.30	0.69	0.40
5	0.99	1.14	0.29	0.39
6	1.51	1.40	0.17	0.40
7	0.52	1.39	0.37	0.46
8	-0.07	1.15	0.52	0.41
9	0.55	1.58	0.35	0.49
10	-0.42	1.01	0.59	0.37
11	1.45	1.26	0.19	0.38
12	1.01	1.09	0.29	0.38
13	0.60	1.33	0.35	0.45
14	1.37	1.54	0.18	0.43
15	-0.33	1.33	0.58	0.44
16	0.99	1.30	0.27	0.43
17	0.62	1.43	0.35	0.46
18	0.52	1.20	0.38	0.42
19	1.10	1.56	0.23	0.46

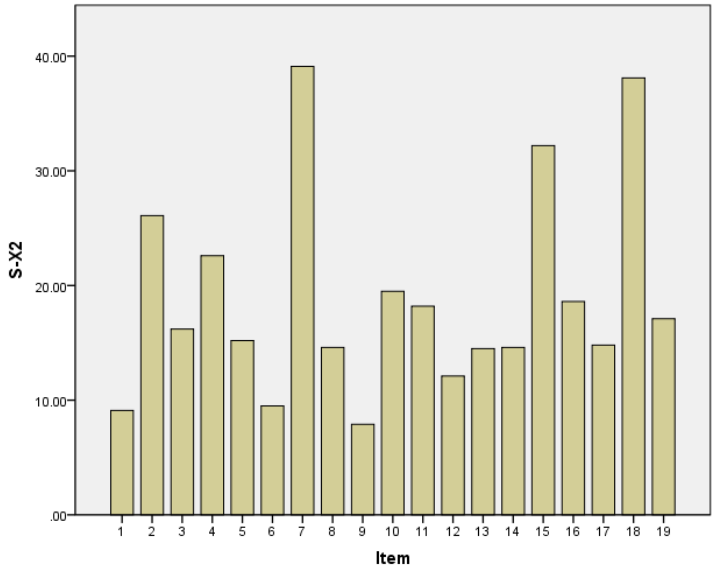
Because a 2PL model was used, ICCs for the SAS item were dependent on both the *b* and *a*-parameters. The ICCs provided a visual of the item parameters found in Table 4, specifically of how the items function across the θ distribution. Appendix C contains all ICCs.

To further evaluate how well the 2PL model fits the data, item-level χ^2 fit statistics (Orlando & Thissen, 2000) were calculated through FlexMIRT. Only items 7, 15, and 18, shown in Table 5, were significant at $p < 0.01$ and flagged for possible misfit. Figure 1 displays all item fit statistics, as does Appendix D.

Table 5. Items Flagged for Possible Misfit

Item	S-χ^2	p
7	39.1	< 0.01
15	32.2	< 0.01
18	38.1	< 0.01

Figure 1. Item Fit Statistics



The last step to evaluate the degree to which the 2PL model explained the data was to analyze whether the assumption of local independence holds. The JSI index produced through FlexMIRT output, as well as results from DIMTEST, were used to analyze whether items exhibited local independence. Table 6 displays the JSI output from FlexMIRT.

The table displays the JSI index for each item pair when one item is removed. The item removed is listed as the column, and the remaining item, the item examined, is listed as the row. For example, the first value in the fourth row indicates that when item 1 is

removed, the JSI value for item 4 is 0.4. This means that by removing item 1, the slope of item 4 decreased. If the JSI value was negative, then by removing an item, the slope of the item increased. The index is interpreted similarly to a standard error, in that the JSI is the standard error of the slope of one item after removing another item.

Bolded and shaded values in the table indicate JSI values that are much higher than other values. As indicated by Edwards, Houts, and Cai (2017), there is currently no metric threshold for deciding when item pairs are locally dependent; however, item pairs that have JSI values that appear as outliers when compared to the other JSI values are generally considered as possible violations of local independence. Thus, a greater than one rule of thumb was used to flag item pairs.

According to the JSI indices calculated in FlexMIRT, several item pairs indicated a possible violation of the local independence assumption. Specifically, the following item pairs may exhibit local dependence: items 3 and 17; items 8 and 9; items 9 and 19; items 11 and 15; 13 and 12; 13 and 16; 14 and 16; 15 and 10; 15 and 11; 16 and 10; 16 and 13; 16 and 14; and 17 and 3. Notably, many of these pairs are items 10-16, which measure contradictory feelings on love.

Table 6. JSI Indices

Item	Item Removed																		
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	--	0.5	0.5	0.5	0.2	-0.1	0.0	-0.2	-0.3	0.0	-0.4	0.0	-0.1	0.3	-0.3	-0.2	0.1	0.2	0.0
2	0.4	--	0.5	0.2	0.5	0.1	0.2	0.4	0.5	-0.2	-0.3	-0.3	-0.4	-0.3	-0.1	-0.7	-0.1	0.0	0.1
3	0.3	0.4	--	0.1	0.0	-0.2	0.4	0.2	0.6	-0.6	0.0	-0.3	-0.7	-0.8	0.3	-1.0	**1.2**	0.4	0.5
4	0.4	0.1	-0.2	--	0.4	-0.1	0.8	0.4	-0.2	-0.1	-0.5	0.1	-0.5	0.0	-0.5	-0.2	-0.1	0.1	-0.1
5	0.1	0.5	0.0	0.5	--	0.1	0.6	0.9	0.3	-0.2	-0.5	-0.4	-0.4	-0.1	-0.6	-0.3	0.0	0.3	-0.2
6	-0.1	0.1	-0.2	-0.1	0.1	--	0.4	-0.2	0.5	0.1	0.4	0.0	0.3	0.1	-0.4	0.2	-0.3	-0.2	0.1
7	-0.1	0.1	0.4	0.8	0.5	0.2	--	0.2	0.1	-0.1	-0.2	-0.5	-0.4	-0.1	-0.4	-0.4	0.4	0.3	-0.3
8	-0.2	0.4	0.2	0.4	0.8	-0.2	0.1	--	**1.4**	-0.9	-0.6	-0.3	-0.9	-0.8	-0.5	-0.8	0.2	0.8	0.9
9	-0.3	0.5	0.6	-0.1	0.3	0.3	0.1	**1.3**	--	-0.7	-0.3	-0.3	-0.6	-0.7	-0.3	-0.8	0.2	0.3	1.0
10	-0.5	-0.2	-0.8	-0.2	-0.3	0.1	-0.2	-1.0	-0.9	--	0.6	0.1	1.0	0.8	**1.1**	**1.1**	-0.6	-0.7	-0.4
11	-0.4	-0.4	0.0	-0.5	-0.6	0.3	-0.2	-0.6	-0.4	0.6	--	0.3	0.8	0.5	**1.6**	0.5	-0.3	-0.5	-0.1
12	-0.1	-0.4	-0.4	0.2	-0.4	-0.1	-0.6	-0.3	-0.4	0.1	0.3	--	**1.2**	0.8	0.4	0.6	-0.4	0.0	-0.1
13	-0.2	-0.6	-0.8	-0.4	-0.4	0.2	-0.4	-0.8	-0.7	0.9	0.6	1.0	--	0.8	0.6	**1.4**	-0.5	-0.4	-0.2
14	0.2	-0.4	-0.9	0.0	-0.1	0.0	-0.1	-0.8	-0.8	0.7	0.4	0.7	0.9	--	0.5	**1.6**	-0.4	-0.5	-0.7
15	-0.4	-0.2	0.2	-0.4	-0.6	-0.4	-0.5	-0.5	-0.4	1.0	**1.2**	0.2	0.5	0.3	--	0.5	-0.1	-0.4	-0.1
16	-0.1	-0.8	-1.2	-0.3	-0.2	0.2	-0.4	-0.8	-0.9	1.0	0.5	0.6	**1.6**	**1.7**	0.6	--	-0.6	-0.4	-0.6
17	0.1	0.1	**1.3**	0.0	0.0	-0.2	0.4	0.2	0.3	-0.5	-0.2	-0.3	-0.5	-0.4	0.0	-0.6	--	1.0	-0.2
18	0.1	-0.2	0.5	0.3	0.3	-0.3	0.3	0.8	0.3	-0.5	-0.5	0.0	-0.4	-0.5	-0.3	-0.4	1.0	--	0.2
19	0.1	0.1	0.6	-0.1	-0.2	0.1	-0.2	0.9	**1.2**	-0.4	0.0	0.0	-0.1	-0.6	0.0	-0.6	-0.1	0.3	--

As noted previously, the DIMTEST analysis revealed that the SAS data may not be unidimensional. In comparing the items chosen for the AT group in the DIMTEST analysis with the JSI indices, a pattern emerges. Items 12, 13, 14, and 16 were determined to be dimensionally similar to each other, and different from all other items in the EFA. These item pairs also exhibited local dependence as exhibited through the JSI indices.

Which SAS Items Provide the Most Information? Does the Most Information Occur Near High Scores on the SAS?

To evaluate which items provide the most information across the θ scale, and ultimately, where the most information occurs, IIFs and the TIF were generated through R (R Core Team, 2017), and Cronbach's α was calculated in SPSS Version 19.

Item Information

In evaluating the items that prove the most information, particularly near the higher point on the scale, the researcher found that items 2, 3, 9, 14, and 19 provided the most information where higher scores occur. These IIFs are displayed in Figures 1-5. The remaining IIFs are provided in Appendix E.

Figure 2. Item Information Function for Item 2

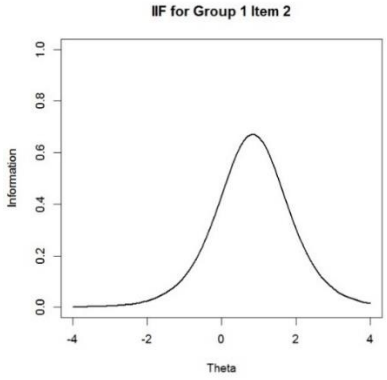


Figure 3. Item Information Function for Item 3

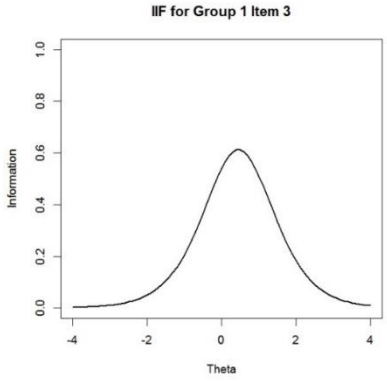


Figure 4. Item Information Function for Item 9

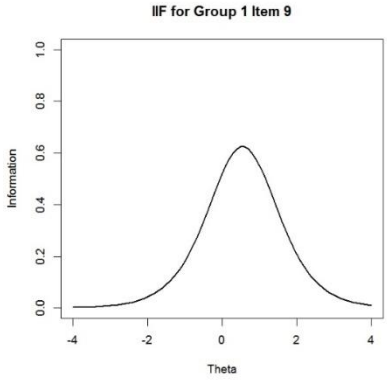


Figure 5. Item Information Function for Item 14

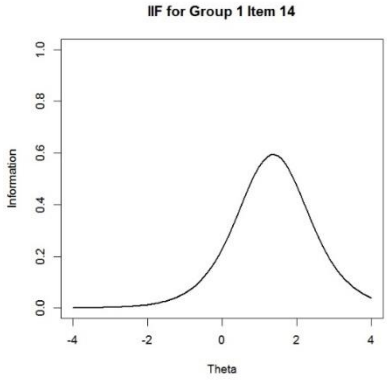
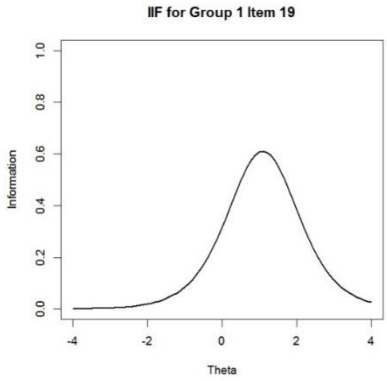


Figure 6. Item Information Function for Item 19



Test Information

The SAS has a reliability of $\alpha = 0.84$, which is consistent with previous studies (Kwapil et al., 2002). Test information is highest at $\theta = 0.8$, occurring on the higher end of the θ scale. Though the SAS does not have a cut score given the nature of the ambivalence construct, possessing more ambivalence indicates a higher risk for developing schizotypy. As such, one would hope that the most information on the SAS should occur at the higher end of the θ scale. Indeed, information is higher end of the

SAS θ scale, and the corresponding standard error of measurement (SEM) is lowest at the higher end of the SAS θ scale. The information and standard error at various θ are shown in Table 7; The TIF is found in Figure 7, and the SEM is found in Figure 8.

Table 7. Information at Various θ Points

θ	-2.8	-2.3	-2.0	-1.6	-1.2	-0.8	-0.4	0.0	0.4	0.8	1.2	1.6	2.0	2.4	2.8
Information	1.48	1.76	2.19	2.79	3.62	4.68	5.94	7.25	8.30	8.71	8.26	7.10	5.62	4.22	3.10
SE	0.82	0.75	0.68	0.60	0.53	0.46	0.41	0.37	0.35	0.34	0.35	0.38	0.42	0.49	0.57

Figure 7. Test Information Function

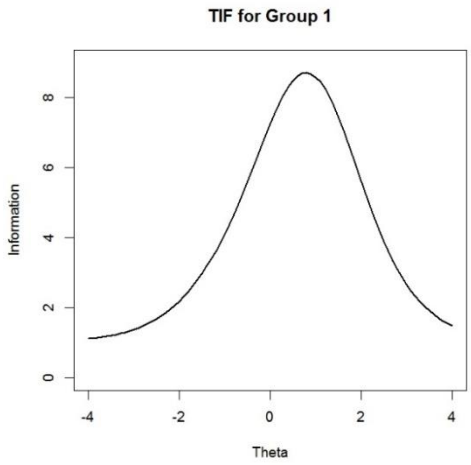
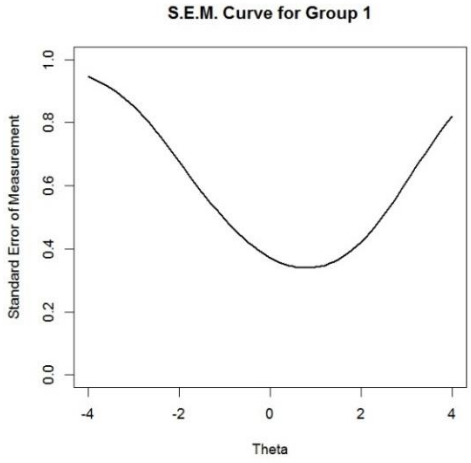


Figure 8. Standard Error of Measurement



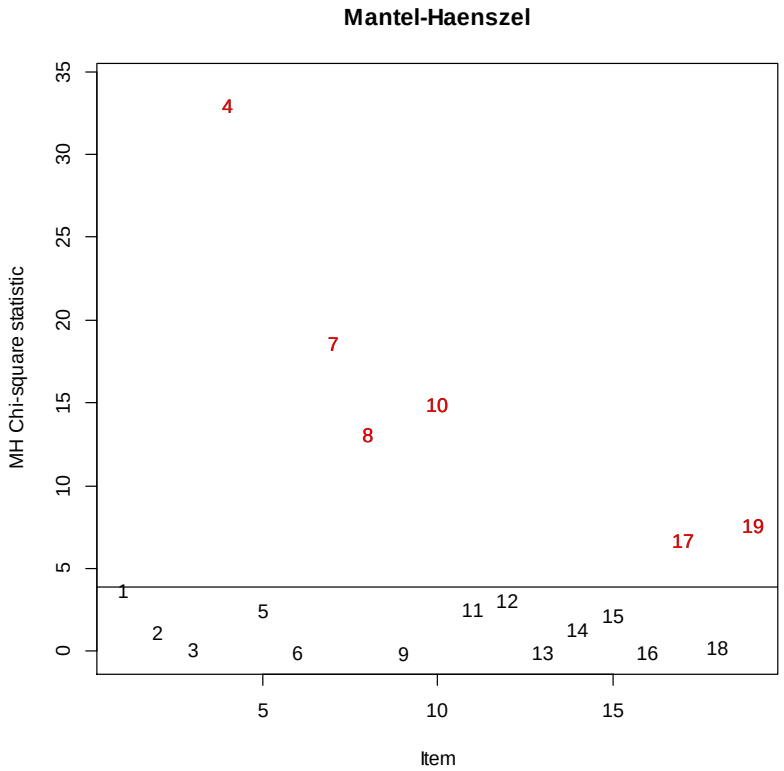
Do any Items Exhibit DIF?

DIF was investigated to uncover whether there were any differences in responses by sex, where males were the focal group and females were the reference group. The researcher used MH DIF and SIBTEST (Shealy & Stout, 1993) to assess whether any

items possess DIF, where potentially problematic items are those items that have DIF values that fall into the moderate or large/significant categories.

Results from the MH DIF analyses revealed that only one of the 19 items exhibited moderate DIF (B), item 4. Item 4 exhibited moderate DIF against the reference group, males, MH D-DIF = 1.01. That is, when compared to males at the same θ level, females more often endorsed item 4. Items 7, 8, 10, 17, and 19 were flagged as exhibiting negligible DIF (A). Figure 9 displays the MH χ^2 statistic for each SAS item.

Figure 9. Mantel-Haenszel χ^2 Statistics



SIBTEST results confirmed the results from the MH DIF analysis – only item 4 exhibited moderate DIF (B) in favor of females, $\beta_{uni} = -0.069$. Like MH DIF results, items 7, 8, 10, were also flagged as items exhibiting negligible DIF in favor of females; item 19 was also flagged as exhibiting negligible DIF in favor of males. Interestingly, SIBTEST flagged items 11, 12, and 15 as exhibiting negligible DIF in favor of males. Overall, results from the MH DIF and SIBTEST indicate that only one item, item 4, should be further investigated for DIF effects.

CHAPTER V

DISCUSSION

The purpose of this research was not only to illustrate the IRT "tool chest" at one's disposal, but to illustrate the possibilities of its use in the field of psychology and the importance of using multiple methods of analysis before drawing conclusions. Given the limited research on the SAS and its structure, the dimensionality of the SAS was first investigated. This analysis ensured that an appropriate IRT model was used, and that the internal structure of the SAS was as Raulin (1986) intended. Next, the data were modeled via three IRT models, and the fit indices of those models were compared to identify the model that best fit the data. From there, the item parameters and characteristics across IRT and CTT were compared to identify item quality and to provide some information regarding model utility, or how well the model chosen helped to meaningfully explain the data. Additional statistics, such as individual item fit and local independence, were evaluated to assess the relationship of items and to flag any problematic issues. Finally, DIF analyses were performed to examine whether any differences existed between men and women in their performance on the items that could not be accounted for by the ambivalence construct. For all analyses, the relevant *Standards* (AERA, APA, & NCME, 2014) were cited, and evidence for the *Standards* was confirmed; in a few instances, the evidence or lack thereof highlighted areas for improvements to the SAS.

This chapter is organized by the focal analyses and addresses findings from each analysis, as well as the consistencies or differences across methods used. When possible, references from relevant literature are cited, especially when results confirm or refute existing research on ambivalence or the SAS. Finally, each section discusses the implications of results as they relate to the *Standards* (AERA, APA, & NCME, 2014). The chapter concludes by identifying some of the limitations of the study, as well as some recommendation for future research on the SAS and on the *Standards*.

Dimensionality of the SAS

Investigation into the dimensionality of the SAS yielded results that confirmed the SAS as essential unidimensional. This conclusion was reached by using two rules that have been established as successful in other research into essential unidimensionality (Slocum-Gori & Zumbo, 2011), the 4:1 and the 3:1 rule. The dimension extracted appears to capture the contradictory thoughts or feelings characteristic of ambivalence as described by Bleuler (1911, 1950). However, the 4:1 rule was not explicitly met, so both DIMTEST and a PCA analysis on a tetrachoric correlation matrix were conducted. Results from the PCA confirmed essential unidimensionality, however, results from the DIMTEST analysis suggest that the underlying structure of the SAS data may not be unidimensional. Results from the DIMTEST analysis are discussed further in the local independence summary.

These results conflict with some of the limited available research. MacAuley et al. (2014) found three different factors through EFA. However, one must consider the sample size, population, and item types. While the sample used in the present study

included more than 7,000 respondents, the MacAuley et al. (2014) study included only 334 participants. And, while MacAuley et al. (2014) removed participants whose ages were outside of the range of 18-25 years, the current study did not exclude data based on participants' ages. Finally, MacAuley et al. (2014) transformed the dichotomous SAS items into Likert-style items, which precludes a direct comparison of results.

The difference in response formats or item types may have impacted EFA results. In some cases, Response scales with more options have resulted in more dimensions (Barendse, Oort, & Timmerman, 2015), but not always. Clark and Watson (1995) argue the advantages and disadvantages of both types, noting that Likert-style items "may reduce validity if respondents are unable to make the more subtle distinctions that are required" (p. 313), but that Likert-style items allow for people to choose from a range. Notably, Clark and Watson (1995) argue that the item types do not produce different factor structures, citing results from an evaluation of a neuroticism scale (Watson, Clark, & Harkness, 1994). However, both item types under the SAS should be piloted with similar groups and then compared to confirm that the underlying factor structure is consistent.

Differences in the sample sizes between the two studies is also worth noting. In considering what constitutes an adequate sample size, several rules of thumb have been proposed, including a sample of at least 100, 200, and upwards of 500. Other rules have stipulated that the sample size is dependent on the ratio of variables to factors. Pearson and Mundfrom (2010) investigated sample size as it relates to EFA and confirmed that

dichotomous data need larger samples, but that a sample size of 100 was adequate for one- or two-factor models, and three-factor model if communalities were high.

Hattie (1985) noted limitations to most decision rules for establishing unidimensionality, several of which dealt with eigenvalue rules. For example, had the rule of counting any eigenvalue greater than one (Kaiser, 1960) been used, one would have concluded that the SAS had two factors, regardless of how little variance the second factor accounted for. Assessing dimensionality is often more art than science, and the decision rules for establishing dimensionality can often contradict each other. Literature has shown that there are strengths and weaknesses in each method depending on the data and type of measure.

As previously noted, the *Standards* specify that "the extent to which item interrelationships bear out the presumptions of the framework would be relevant to validity" (AERA, APA, & NCME, 2014, p. 16). Evidence of item interrelationships, or the internal structure of the of the SAS, can provide one piece of the validity argument or one method of establishing construct validity (Slocum-Gori & Zumbo, 2011, p. 443). The dimensionality analysis of the SAS provided evidence for essential unidimensionality of the SAS, but as other analyses indicated, such as the results from local independence, further research regarding the internal structure is warranted. Perfect unidimensionality is all but improbable, and thus the discussion becomes about the degree to which the departure from unidimensionality requires one to account for multiple dimensions via subscores. This is not necessarily a weakness, as "the validation process never ends, and there is always additional information that can be gathered to more fully understand a test

and the inferences that can be drawn from it" (AERA, APA, & NCME, 2014, pp. 21-22). Thus, while the analysis did not provide a definitive answer on the dimensionality of the SAS, the analysis provided adequate evidence for *Standard 1.13*.

Model Fit

After establishing that the SAS is essential unidimensional, the data were modeled via three unidimensional IRT models, and the best fitting IRT model according to global fit indices was chosen. The RMSEA indicated no difference in fit across the three models. However, the AIC indicated that the 2PL model best fit the data. As expected, the 3PL model was not chosen, as the guessing parameter (lower asymptote) is less prevalent in psychological assessments than in educational assessments. That is, it is unlikely that a person would guess for a response on a self-report scale. The 2PL model was expected to fit the data best given the purpose of the SAS measuring the ambivalence construct. Specifically, one would expect that some items would be written to be more discriminating than others at certain points on the scale. And historically, the 2PL model is a conceptually consistent result in psychological measurement (Edwards, Houts, & Cai, 2017, p. 2).

Ultimately, the model fit analysis provided evidence for which model fit the data best, evidence for *Standard 1.13* (AERA, APA, & NCME, 2014) or evidence regarding the internal structure of the SAS, and evidence for *Standard 4.10*, which indicates that documentation is required regarding the psychometric properties of items, that a rationale for the sample used should be provided, and that evidence be provided of model fit.

While the analysis of model fit only partially address *Standard 4.10*, it is necessary step before evaluating item parameters.

Model Utility

In building upon model fit analyses, SAS items were analyzed via 2PL item parameters, CTT statistics, item fit statistics, ICCs, and local independence. The results indicated that the IRT and CTT item-level statistics were comparable, with items 4, 8, 10, and 15 endorsed at the highest rate and items 1, 6, 11, and 14 endorsed at the lowest rate. Items 10 and 12 were the least discriminating; item 10 asked respondents about love and hate, while item 12 asked respondents about their feelings about getting close to people and their faults.

Item fit statistics revealed three items that exhibited poor fit, including items 7 ("I always seem to be the most unsure of myself at the same time that I am most confident of myself"), 15 ("My experiences with love have always been muddled with great frustrations"), and 18 ("I usually experience doubt when I have accomplished something that I have worked on for a long time"). The limited number of items exhibiting possible misfit is impressive, and why these items exhibited issues with fit is puzzling. The item parameters for these items did not stand out; in fact, the a-parameters for these items were high. Furthermore, these items did not exhibit local dependence. Future research may include investigation into these items and their fit indices across samples.

Evaluation of local independence through the JSI index revealed several items that exhibited local dependence. Items 13, 14, 15, and 16 especially all had high JSI indices for each pair. Notably, these items dealt with feelings towards another person,

specifically the juxtaposition of loving someone and a competing feeling such as hate, anger, or frustration. These item pairs violate the assumption of local independence, which would indicate that the 2PL model may not fully explain the data. Notably, results from the DIMTEST analysis confirm that items 12, 13, 14, and 16 may be dimensionally distinct from the other items while being like each other. Given results from the eigenvalue ratios, these items may measure another construct.

One explanation for this violation is that these items could also be exhibiting a different type of local dependence known as surface local dependence (SLD; Chen & Thissen, 1997). SLD occurs "when respondents answer two questions identically because the items are similar, either in content or location on the instrument" (Edwards, Houts, & Cai, 2017, p. 2). Items 12 through 16 are similar in content – each item addresses one's views of love, which may provide respondents with difficulty in differentiating their responses. These items are also similar in location on the SAS, which may encourage respondents to answer these items similarly. More research is needed into the SAS and SLD, and how the JSI index functions with SLD.

However, another explanation is that these items could simply be measuring multiple constructs. DIMTEST results suggested that items 12, 13, 14, and 16 were dimensionally similar to one another. Despite differences in methodological approaches and item types, the factor underlying these items could align with the interpersonal dimension identified by MacAuley et al. (2014). However, the third factor identified by MacAuley et al. (2014) is absent from the current results.

Results from the analysis of the item characteristics revealed that in general, the model provided some utility in explaining the data, but it also presented some limitations as indicated by issues with local independence. Conducting these analyses provides evidence in support of *Standard 4.10*, which requires that one document the psychometric properties of a scale, though results from the test of local independence did not fully support *Standard 1.13* (AERA, APA, & NCME, 2014). While the violation of local dependence does not warrant removing these items, especially given the quality of other item characteristics, further research might investigate whether distributing items 12 through 16 across the SAS scale would reveal if the location of these items is resulting in SLD.

Information

The construct of ambivalence is viewed as fluid by psychologists; thus, the SAS does not have a definitive cut at which one would determine that a person does or does not possess ambivalence. Though the SAS does not have a cut score, higher scores on the SAS are intended to indicate that one possesses more of the ambivalence construct, and thus may be predisposed to later develop schizotypy or schizophrenia. Thus, one would hope that the most information or precision in SAS scores occurs at the higher end of the ambivalence continuum. In evaluating information at the item level, the IIFs indicated that most items target the higher end of the θ scale. Exceptions included items 4, 8, 10, and 15; these items provided the most information at or below 0 on the θ scale.

Test-level results regarding reliability indicated that scores from the SAS are reliable, as Cronbach's $\alpha = 0.84$; this statistic is consistent with previous research

(Kwapil et al., 2002). However, IRT information provided more precision, and identified that the most information occurred at the higher end of the θ scale, specifically where θ = 0.8. This result is reasonable, given that higher scores on the SAS are intended to indicate higher levels of ambivalence, which can be an early marker for later development of schizotypy or schizophrenia. Thus, the most confidence should be in SAS scores that are on the higher end of the latent trait scale. However, information appears to decrease and error appears to increase at the highest end of the scale, indicating that there is less information for the highest scores on the SAS. However, in general, the TIF and Cronbach's α provide evidence for *Standards 2.0* and *2.3*, and documentation of evidence for *Standard 7.12* (AERA, APA, & NCME, 2014). However, if the SAS were to ever introduce a diagnostic cut score, the SAS might require additional items towards the higher end of the θ scale to ensure greater confidence near the cut.

DIF

Results from the MH DIF and SIBTEST analyses showed that only item 4 exhibited moderate DIF. This item, "Very often when I feel like doing something, at the same time I don't feel like doing it," indicated that males were less likely to endorse the item than females. That is, female participants were more likely to endorse the item than males *after* males and females were matched with respect to the underlying construct. Excluding item 4, results from the DIF analysis provide evidence that psychologists can use the SAS without concerns of construct-irrelevant differences between males and females. Given the quality of the item-level characteristics, the researcher suggests keeping item 4. The results also provided in support of *Standard 4.10* (AERA, APA, &

NCME, 2014) which requires that documentation of analyses such as DIF be provided; of *Standard 3.0*, which requires that construct-irrelevant variance be minimized; and *Standard 3.1*, which requires that test developers take steps to understand differences in subgroups and to maximize access for all.

Implications of Results

Results from this study shows that the SAS is a relatively robust measure of the ambivalence construct, and that in general, most items appear to provide ample information near the higher end of the trait scale. The study also provided evidence that the SAS can be used across gender subgroups without concerns about DIF. The results also emphasize that IRT has utility in psychology, and that psychologists should more often turn to IRT when evaluating measures with diagnostic outcomes. Many of the IRT tools provide greater information than CTT counterparts. Both the IIFs and TIFs were more advantageous than Cronbach's α because the IIFs and TIF provided precision data at various points on the θ scale. Without IRT, psychologists would simply know that the SAS is reliable, not where the SAS provides the most information regarding scores. The latter is of utmost importance for the SAS, for which high scores indicates a predisposition to later development of schizotypy.

The study was also an exercise in how the *Standards* (AERA, APA, & NCME, 2014) can be used guiding the evaluation of the quality of psychological assessments. By considering how each of the analyses provide evidence for the *Standards*, a psychologist can articulate an argument for and provide technical documentation of the quality of a scale or assessment. Documentation of the utility and quality of the SAS is strengthened

by aligning analyses and findings to *Standards* widely accepted by the field. Furthermore, the *Standards* allow one to pinpoint weaknesses in the scale or analyses, allowing for future improvements. While the *Standards* do not have a defined set of rules for scales used to monitor rather than diagnose, future iterations of the *Standards* may provide clearer guidance on formative assessments or scales used for purposes other than diagnostics.

Finally, this study illustrates the importance of an integrative approach to analyses. Multiple pieces of evidence are needed to illustrate that one can use a scale for a specific purpose, and that the scores resulting from that scale are valid. As demonstrated, existing methods for establishing dimensionality can contradict one another, and using only one method may result in inaccurate conclusions. Similarly, using multiple methods for establishing DIF provided stronger evidence for the lack of DIF items. In any evaluation of a psychological scale or educational assessment, the goal should be to create a coherent argument by using and comparing multiple methods of analysis.

Limitations and Directions for Future Research

The current study is not without limitations. While the sample represented university students whose ages most likely fell within the range of ages for which one develops schizotypy or schizophrenia, the sample was not a clinical sample, and it is possible that some participants were outside of the 18-25 years age range.

Future research should include further analysis of the internal structure of the SAS. Using the eigenvalue ratio rules allowed for the claim of essential

unidimensionality. However, given findings from the DIMTEST procedure and the violation of local independence by several item pairs in the current study as shown by the JSI indices, dimensionality and local independence should be investigated further. Some items on the SAS may measure multiple factors, or items phrased similarly may pose issues for respondents in discriminating their responses. An investigation into dimensionality and local independence – including SLD – would provide more confidence regarding the internal structure of the SAS, and would strengthen the existing evidence in support of *Standard 1.13* (AERA, APA, & NCME, 2014). As a corollary, further investigation into different response formats may yield more information regarding the response format best suited for the SAS.

Future research may warrant an evaluation of how item 4 functions for various subgroups. Currently, the odds of females responding to this item is greater than males. Researchers should consider why females are more likely to endorse the item than males with the same level of the underlying trait, and whether the item could be rewritten in a way that does not favor one subgroup. Investigation of DIF should also include other subgroups such as those related to ethnicity. Given preliminary differences identified in Mann et al. (2008), evaluating differences between Caucasians and African-American warrants further study. Investigating DIF for other subgroups would provide greater strength of evidence for the *Standards* (AERA, APA, & NCME, 2014), particularly for *Standards 3.0* and *3.1*.

Lastly, future research may warrant evaluating the SAS in terms of other *Standards* (AERA, APA, & NCME, 2014). The scope of this study was limited to

research questions pertaining to the internal structure of the SAS and the psychometric properties of the SAS. Thus, other relevant *Standards* were excluded. For example, while reliability has been established for the SAS, these reliability studies have not included connections to the reliability *Standards*. Future work could also integrate *Standard 1.2*, or the intended use of scores from the SAS. Though the intended interpretation of the SAS has been described in previous research, pulling in the relevant *Standards* provides greater confidence for those using and interpreting results from the SAS.

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transaction on Automatic Control, 19*, 716-723.

Allen, M. J, & Yen, W. M. (1979). *Introduction to measurement theory*. Long Grove, IL: Waveland Press.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing (2014). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

American Psychiatric Association (2013). Diagnostic and statistical manual of mental disorders (5th ed.). Washington, DC: American Psychiatric Association.

Apinyapibal, S., Lawthong, N., & Kanjanawasee, S. (2015). A comparative analysis on the efficacy of differential item functioning detection for dichotomously scored items among logistic regression, SIBTEST, and raschtree methods. *Procedia – Social and Behavioral Sciences, 191*, 21-25.

Barendse, M.T., Oort, F.J., & Timmerman, M. E. (2015). Using exploratory factor analysis to determine the dimensionality of discrete responses. *Structural Equation Modeling: A Multidisciplinary Journal, 22*, 87-101. doi: 10.1080/10705511.2014.934850

Birnbaum, A. (1968). *Some latent trait models and their use in inferring an examinee's ability*. In Lord, F. M. and Novick, M. R. (Eds.), Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Bleuler, E. P. (1950). *Dementia praecox of the group of schizophrenias*. (J. Zinkin, Trans.). New York: International Universities Press (original work published 1911).

Bock, R. D., & Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika, 35,* 179-197.

Cai, L. (2013). FlexMIRT version 2: Flexible multilevel multidimensional item analysis and test scoring [Computer software]. Chapel Hill, NC: Vector Psychometric Group.

Canel-Çinarbas, D., Cui, Y., & Lauridsen, E. (2011). Cross-cultural validation of the Beck Depression Inventory-II across U.S. and Turkish samples. *Measurement and Evaluation in Counseling and Development, 44*, 77-91.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R Environment. Journal of Statistical Software, 48, 1-29.

Chapman, J. P., Chapman, L. J., & Kwapil, T. R. (1994). Does the Eysenck psychoticism scale predict psychosis? A ten year longitudinal study. *Personality and Individual Differences, 17,* 369-375.

Chen, W., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22,* 265-289.

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment, 7*, 309-319.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297-334.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281-302.

Davenport Jr., E. C., Davison, M. L., Liou, P.Y., & Love, Q. U. (2015). Reliability, dimensionality, and internal consistency as defined by Cronbach: Distinct albeit related concepts. *Educational Measurement: Issues and Practice: 34,* 4-9.

Di Nardo, P. A., Brown, T. A., & Barlow, D. H (1994). *Anxiety disorder interview schedule for DSM-IV: Lifetime version (ADIS-IV-L).* Oxford University Press: NY.

Dorans, N., & Holland, P. (1993). DIF detection and Description: Mantel-Haenszel and Standardization. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35-66). Hillsdale, NJ: Erlbaum.

Edwards, M.C., Houts, C.R., & Cai, L. (2017). A diagnostic procedure to detect departures from local independence in item response theory models. *Psychological Methods*. Advance online publication. doi: 10.1037/met0000121.

Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment, 8,* 341-349.

Embretson, S.E., & Reise, S.P. (2000). *Item response theory for psychologists*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Emmert-Aronson, B.O., Moore, M.T., & Brown, T.A. (2014). Differential item functioning of the symptoms of major depression by race and ethnicity: An item response theory analysis. *Journal of Psychopathology and Behavioral Assessment, 36,* 424-431.

Freud, S. (1913). Totem and taboo. *Standard Edition, 13,* 1-161.

Freud, S. (1917). Mourning and melancholia. *Standard Addition, 14,* 243-258.

Green, B. F. (1956). A method of scalogram analysis using summary statistics. *Psychometrika, 21,* 79-88.

Green, S. B., Lissitz, R. W., & Mulaik, S. A. (1977). Limitations of coefficient alpha as an index of test unidimensionality. *Educational and Psychological Measurement, 37,* 827-838.

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review, 80,* 139-150.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice, 12,* 38-47.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Norwell, MA: Kluwer Academic Publishers.

Hattie, J. (1985). Methodology Review: Assessing Unidimensionality of Tests and Items. *Applied Psychological Measurement, 9,* 139-164.

Hattie, J., Krakowski, K., Rogers, H. J., & Swaminathan, H. (1996). An assessment of Stout's index of essential unidimensionality. *Applied Psychological Measurement, 20,* 1-14.

Houts, C. R., & Cai, L. (2013). flexMIRT user's manual version 2: Flexible multilevel multidimensional item analysis and test scoring. Chapel Hill, NC: Vector Psychometric Group.

IBM Corp. Released 2010. IBM SPSS Statistics for Windows, Version 19.0. Armonk, NY: IBM Corp.

Joreskog, K. G. (1978). Structural analysis of covariance and correlation matrices. *Psychometrika, 43,* 443-477.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement, 20,* 141-151.

Kane, M. (2012). Validating score interpretations and uses: Messick lecture, Language Testing Research Colloquium, Cambridge, April 2010. *Language Testing, 29, 3-17.*

Kane, M. (2006). Validation. In R. Brennan (Ed.), *Educational Measurement, 4th ed. (p. 17-64).* Westport, CT: American Council on Education and Praeger.

Kang, T., & Chen, T. (2008). Performance of the generalized S-X^2 item fit index for polytomous IRT models. *Journal of Educational Measurement, 45,* 391-406.

Kernberg, O. F. (1984). *Severe Personality Disorders.* New Haven, CT: Yale University Press.

King, L.A., & Emmons, R.A (1990). Conflict over emotional expression: Psychological and physical correlated. *Journal of Personal and Social Psychology, 58*, 864-877.

King, M. W., Street, A. E., Gradus, J. L., Vogt, D. S., & Resick, P. A. (2013). Gender differences in posttraumatic stress symptoms among OEF/OIF veterans: An item response theory analysis. *Journal of Traumatic Stress, 26*, 175-183.

Kuhn, R., & Cahn, C. H. (2004). Eugene Bleuler's concepts of psychopathology. *History of Psychiatry, 15*, 361-366.

Kwapil, T. R., Raulin, M., & Midthun, J. (2000). A ten-year longitudinal study of intense ambivalence as a predictor of risk for psychopathology. *Journal of Nervous and Mental Disease, 188,* 402-408.

Kwapil, T. R., Mann, M. C., & Raulin, M. L. (2002). Psychometric properties and concurrent validity of the schizotypal ambivalence scale. *Journal of Nervous and Mental Disease, 190,* 290-295.

Lenzenweger, M. F. (2006). Schizotypy: An organizing framework for schizophrenia research. *Association of Psychological Science, 15*, 162-166.

Lord, F. (1953). The relation of test score to the trait underlying the test. *Educational and Psychological Measurement, 13*, 517-549.

Lord, (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum Associates.

MacAulay, R. K, Brown, L. S., Minor, K. S., & Cohen, A. S. (2014). Conceptualizing schizotypal ambivalence: Factor structure and its relationships. *Journal of Nervous and Mental Disease, 202,* 793-801.

MaCallum, R. C., & Austin, J. T. (2000). Applications of structural equation modeling in psychological research. *Annual Review of Psychology, 51,* 201-226.

Magis, D., Beland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods, 42,* 847-862.

Mann, M. C., Vaughn, A. G., Barrantes-Vidal, N., Raulin, M. L., & Kwapil, T. R. (2008). Social ambivalence scale as a marker of schizotypy. *Journal of Nervous and Mental Disease, 196,* 399-404. doi: 10.1097/NMD.0b013e3181710900.

Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of item response theory models. *Measurement, 11,* 71-101.

Meads, D. M., & Bentall, R. P. (2008). Rasch analysis and item reduction of the hypomanic personality scale. *Personality and Individual Differences, 44,* 1772-1783.

Meara, K., Robin, F. & Sireci, S. G. (2000). Using multidimensional scaling to assess the dimensionality of dichotomous item data. *Multivariate Behavioral Research, 35,* 229-259.

Meehl, P. E. (1962). Schizotaxia, schizotypy, schizophrenia. *American Psychologist, 17,* 827-838.

Meehl, P. E. (1989). Schizotaxia revisited. *Archives of General Psychiatry, 46,* 935-944.

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H. Braun, (Eds.), *Test Validity (p. 33-45).* Hillsdale, NJ: Lawrence Erlbaum.

Mislevy, J. L., Rupp, A. A., & Harring, J. R. (2012). Detecting local item independence in polytomous adaptive data. *Journal of Educational Measurement, 49,* 127-147.

Nandakumar, R. (1993). Simultaneous DIF amplification and cancellation: Shealy-Stout's test for DIF. *Journal of Educational Measurement, 30,* 293-311.

Nandakumar, R., & Feng, Y. (1994). Testing the Robustness of DIMTEST on nonnormal ability distributions. *Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.*

Nadakumar, R., & Stout, W. (1993). Refinement of Stout's procedure for assessing latent trait unidimensionality. *Journal of Educational Statistics, 18,* 41-68.

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24,* 50-64.

Pearson, R. H., & Mundfrom, D.J. (2010). Recommended sample size for conducting exploratory factor analysis on dichotomous data. *Journal of Modern Applied Statistical Methods, 9,* 359-368.

Plake, B. S., & Wise, L. L. (2014). What is the role and importance of the revised AERA, APA, NCME standards for educational and psychological testing? *Educational Measurement: Issues and Practice, 33,* 4-12.

Raulin, M. L. (1984). Development of a scale to measure intense ambivalence. *Journal of Consulting and Clinical Psychology, 52,* 63-72.

Raulin, M. L. (1986). *Schizotypal Ambivalence Scale.*

Raulin, M. L., & Brenner, V. (1993). Ambivalence. In CG Costello (Ed), *Symptoms of schizophrenia* (pp. 201-226). New York: Wiley.

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Reise, S. P., Ventura, J., Baade, L. E., Green, M. F., Kern, R.S, Nuechterlein, K. H., … Seidman, L. J. (2011). Bifactor and item response theory analyses of interviewer report scales of cognitive impairment in schizophrenia. *Psychological Assessment, 23,* 245-261.

Revelle, W. (2017). psych: Procedures for personality and psychological research. Northwestern University, Evanston, IL. https://CRAN.R-project.org/package=psych Version = 1.7.8.

Samejima, F. (1969). *Estimation of ability using a response pattern of graded scores* (Psychometrika Monograph No. 17). Richmond, VA: Psychometric Society. Retrieved from https://www.psychometricsociety.org/sites/default/files/pdf/MN17.pdf

Santor, D. A., Ascher-Svanum, H. Lindenmayer, J., & Obenchain, R. L. (2007). Item response analysis of the positive and negative syndrome scale. *BMC Psychiatry, 7,* 66-75.

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics, 6,* 461-464.

Shealy, R., & Stout, W. F. (1993). A model-based standardization approach that separates true bias/DIF from group differences and detects test bias/DIF as well as item bias/DIF. *Psychometrika, 58,* 159-194.

Sincoff, J. B. (1990). The psychological characteristics of ambivalent people. *Clinical Psychology Review, 10*, 43-68.

Silverstein, B. (2002). Gender differences in the prevalence of somatic versus pure depression: A replication. *American Journal of Psychiatry, 159*, 1051-1052.

Slocum-Gori, S. L., & Zumbo, B. D. (2011). Assessing the unidimensionality of psychological scales: Using multiple criteria from factor analysis. *Social Indicators Research, 3,* 443-461.

Slocum-Gori, S. L., Zumbo, B. D., Michalos, A. C., & Diener, E. (2009). A note on the dimensionality of quality of life scales: An illustration with the Satisfaction with Life Scale (SWLS). *Social Indicators Research: An Internal Interdisciplinary Journal for Quality of Life Measurement, 92,* 489-496.

Smith, L., & Reise, S. P. (1998). Gender differences on negative affectivity: An IRT study of differential item functioning on the multidimensional personality questionnaire stress reaction scale. *Journal of Personality and Social Psychology, 75*, 1350-1362.

Steiger, J.H. (2007). Understanding the limitations of global fit assessment in structural equation modeling. *Personality and Individual Differences, 42*, 893-98.

Stout, W. F., & Roussos, L. A. (1996). Simulation studies of the effects of sample size and studies item parameters on SIBTEST and Mantel-Haenszel type I error performance. *Journal of Educational Measurement, 33*, 215-230.

Stout, W. (1987). A nonparametric approach for assessing latent trait dimensionality. *Psychometrika, 52,* 589-617.

Thompson, B., & Daniel, L. G. (1996). Factor analytic evidence for the construct validity of scores: A historical overview and some guidelines. *Educational and Psychological Measurement, 56*, 197-208.

Van Dam, N. T., Earlywine, M., & Borders, A. (2010). Measuring mindfulness? An item response theory analysis of the mindfulness attention awareness scale. *Personality and Individual Differences, 49*, 805-810.

Waller, N.G., Compas, B. E., Hollon, S. D., & Beckjord, E. (2005). Measurement of depressive symptoms in women with breast cancer and women with clinical depression: A differential item functioning analysis. *Journal of Clinical Psychology in Medical Settings, 12*, 127-141.

Watson, D., Clark, L.A., & Harkness, A. R. (1994). Structures of personality and their relevance to psychopathology. *Journal of Abnormal Psychology, 103*, 18-31.

William Stout Institute for Measurement. (2006). Nonparametric dimensionality assessment package DIMPACK (Version 1.0) [Computer software]. St. Paul, MN: Assessment Systems Corporation.

Winterstein, B., Ackerman, T., Silvia, P., & Kwapil, T. (2011). Psychometric properties of the Wisconsin schizotypy scales in an undergraduate sample: Classical test theory, item response theory, and differential item functioning. *Journal of Psychopathology and Behavioral Assessment, 33*, 480-490.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*, 187-213.

Zanon, C., Hutz, C S., Yoo, H., & Hambleton, R. K. (2016). An application of item response theory to psychological test development. *Psychology: Research and Review, 29:18*. doi: 10.1186/s41155-016-0040-x.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P.W. Holland & H. Wainer (Eds.), Differential item functioning (pp. 337-347). Hillsdale, NJ: Erlbaum.

Zwick, W., & Velicer, W. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin, 99*, 432-442.

APPENDIX A

SCHIZOTYPAL AMBIVALENCE SCALE

1. Often I feel like I hate even my favorite activities.
2. My thoughts and feelings always seem to be contradictory.
3. My feelings about my own worth as a person are constantly changing back and forth.
4. Very often when I feel like doing something, at the same time I don't feel like doing it.
5. When I am trying to make a decision, it almost feels like I am physically switching from side to side.
6. It's impossible to know how you feel because the people around you are constantly changing.
7. I always seem to be the most unsure of myself at the same time that I am most confident of myself.
8. I always seem to have difficulty deciding what I would like to do.
9. Most people seem to know what they're feeling more easily than I do.
10. Love and hate tend to go together.
11. Love never seems to last very long.
12. The closer I get to people, the more I am annoyed by their faults.
13. Everyone has a lot of hidden resentment toward his loved one.
14. I have noticed that feelings of tenderness often turn into feelings of anger.
15. My experiences with love have always been muddled with great frustrations.

16. I usually find that feelings of hate will interfere when I have grown to love someone.
17. A sense of shame has often interfered with my accepting words of praise from others.
18. I usually experience doubt when I have accomplished something that I have worked on for a long time.
19. I doubt if I can ever be sure exactly what my true interests are.

APPENDIX B

IRB EXEMPT STATUS

IRB ori@approved-senders.uncg.edu via adminliveunc.onmicrosoft.com May 8

to me, irbcorre, T_KWAPIL, JTWILLSE

To: Lauren Fluegge
Ed Research Methodology
Ed Research Methodology

From: UNCG IRB

Date: 5/08/2017

RE: Notice of IRB Exemption
Exemption Category: 4.Existing data, public or deidentified
Study #: 17-0182
Study Title: Psychometric Analysis of the Schizotypal Ambivalence Scale

This submission has been reviewed by the IRB and was determined to be exempt from further review according to the regulatory category cited above under 45 CFR 46.101(b).

Study Description:

The study will investigate the psychometric properties of the Schizotypal Ambivalence Scale (Raulin, 1986) using archival data from a general population. The study will use item response theory and classical test theory to evaluate the scale's properties.

Investigator's Responsibilities

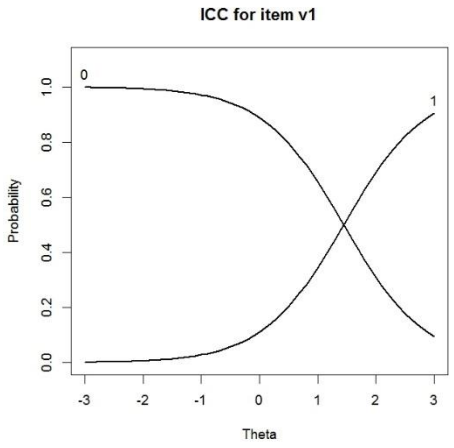
Please be aware that any changes to your protocol must be reviewed by the IRB prior to being implemented. Please utilize the most recent and approved version of your consent form/information sheet when enrolling participants. The IRB will maintain records for this study for three years from the date of the original determination of exempt status.

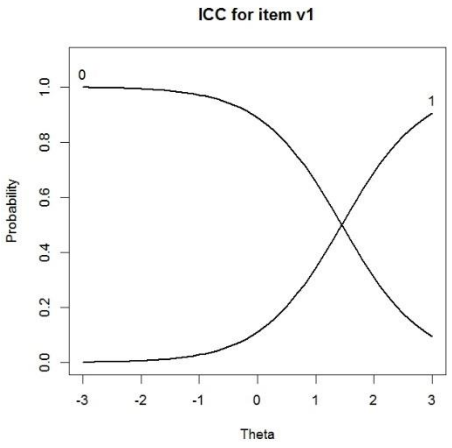
Signed letters, along with stamped copies of consent forms and other recruitment materials will be scanned to you in a separate email. **Stamped consent forms must be used unless the IRB has given you approval to waive this requirement.** Please notify the ORI office immediately if you have an issue with the stamped consents forms.

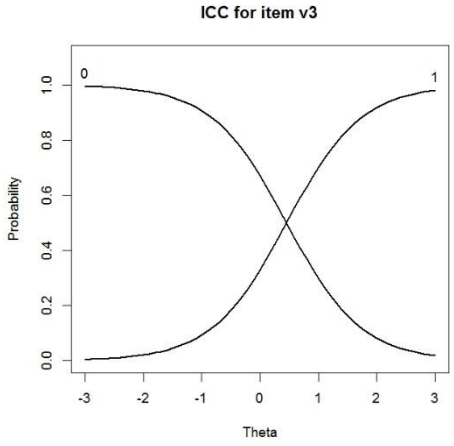
Please be aware that valid human subjects training and signed statements of confidentiality for all members of research team need to be kept on file with the lead investigator. Please note that you will also need to remain in compliance with the university "Access To and Retention of Research Data" Policy which can be found at http://policy.uncg.edu/university-policies/research_data/.

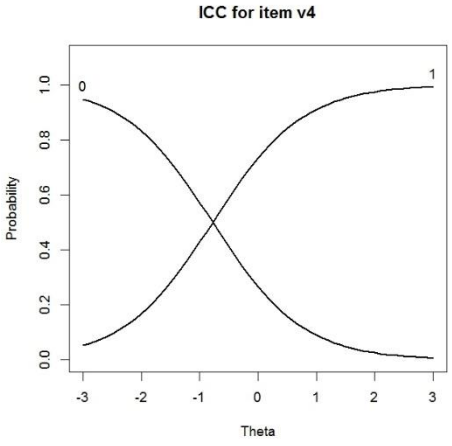
APPENDIX C

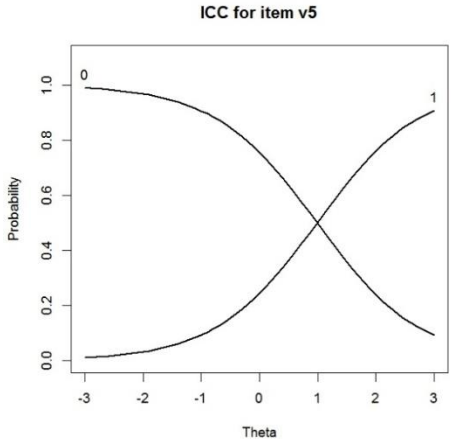
ITEM CHARACTERISTIC CURVES

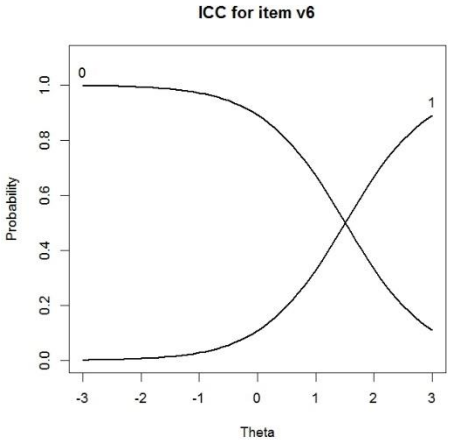


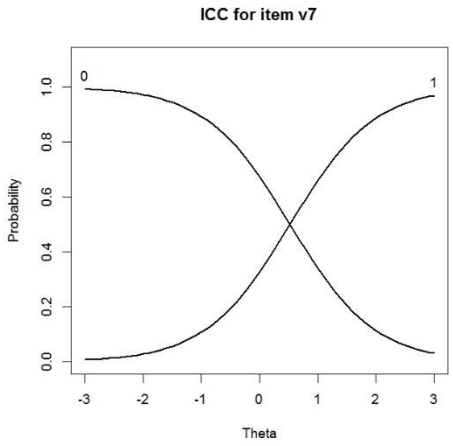


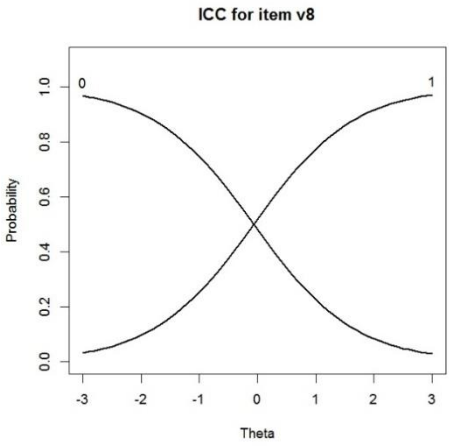


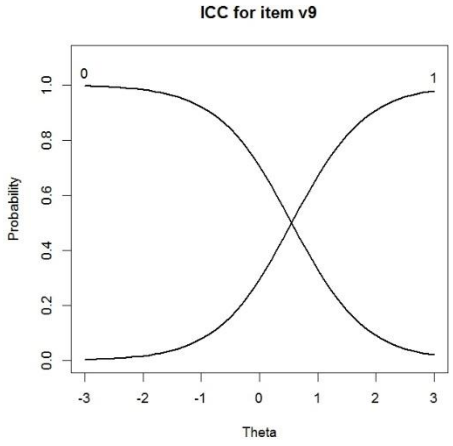


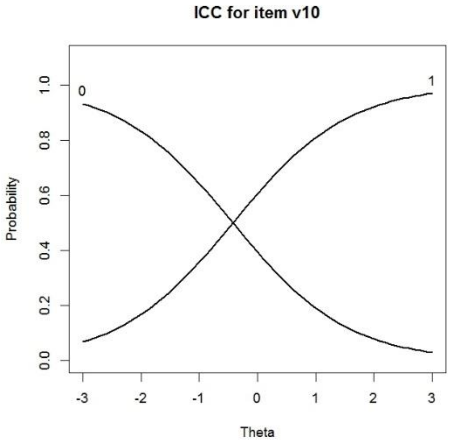

ICC for item v5
Probability
Theta
0
1
0.0
0.2
0.4
0.6
0.8
1.0
-3
-2
-1
0
1
2
3

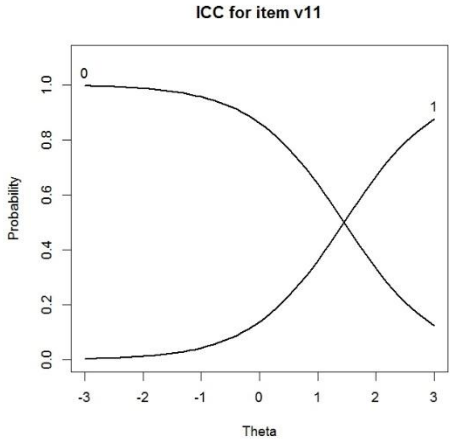

ICC for item v6
Probability
Theta
0
1
0.0
0.2
0.4
0.6
0.8
1.0
-3
-2
-1
0
1
2
3

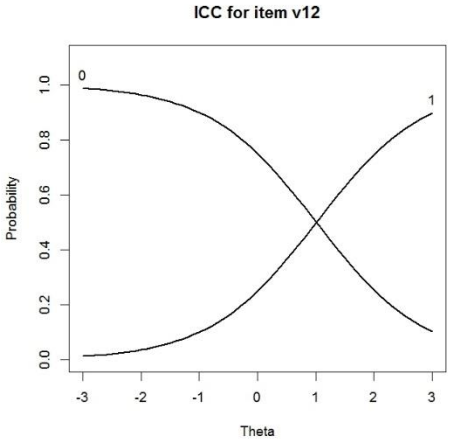

ICC for item v7
Probability
Theta
0
1
0.0
0.2
0.4
0.6
0.8
1.0
-3
-2
-1
0
1
2
3

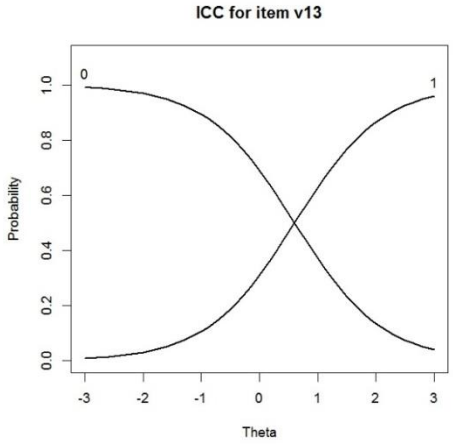

ICC for item v8
Probability
Theta
0
1
0.0
0.2
0.4
0.6
0.8
1.0
-3
-2
-1
0
1
2
3

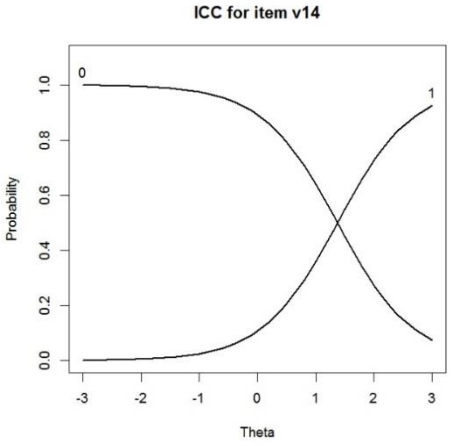

ICC for item v9
Probability
Theta
0
1
0.0
0.2
0.4
0.6
0.8
1.0
-3
-2
-1
0
1
2
3

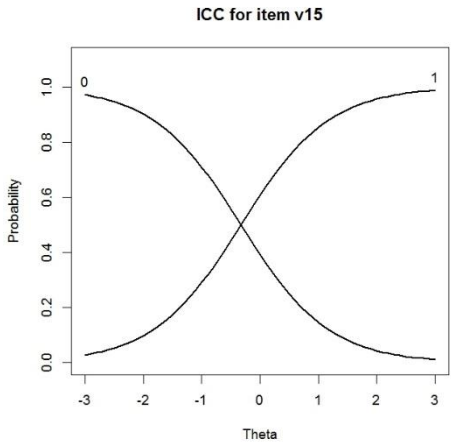

ICC for item v10
Probability
Theta
0
1
0.0
0.2
0.4
0.6
0.8
1.0
-3
-2
-1
0
1
2
3

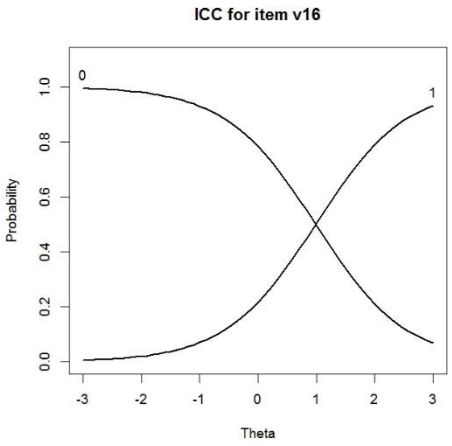

ICC for item v11
Probability
Theta
0
1
0.0
0.2
0.4
0.6
0.8
1.0
-3
-2
-1
0
1
2
3

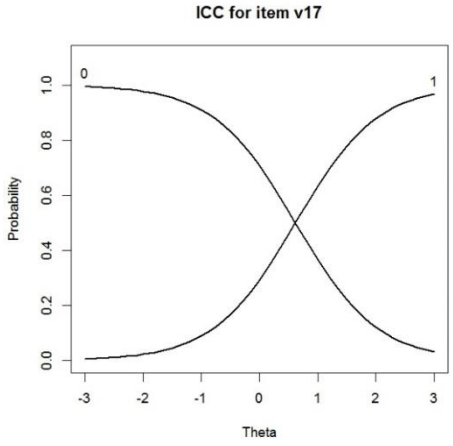

ICC for item v12
Probability
Theta
0
1
0.0
0.2
0.4
0.6
0.8
1.0
-3
-2
-1
0
1
2
3

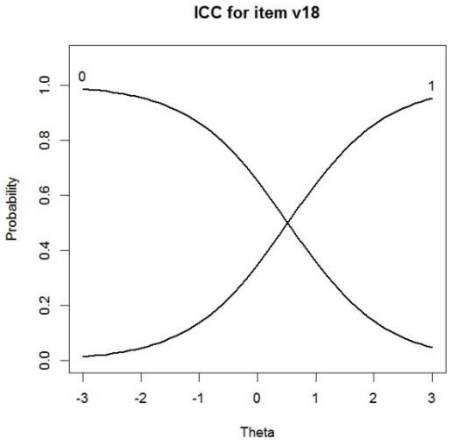

ICC for item v13
Probability
Theta
0
1
0.0
0.2
0.4
0.6
0.8
1.0
-3
-2
-1
0
1
2
3

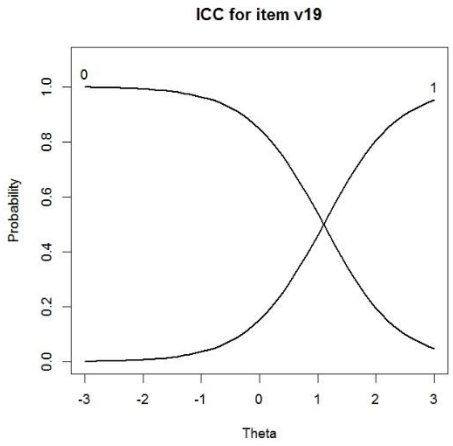

ICC for item v14
Probability
Theta
0
1
0.0
0.2
0.4
0.6
0.8
1.0
-3
-2
-1
0
1
2
3


ICC for item v15
Probability
Theta
0
1
0.0
0.2
0.4
0.6
0.8
1.0
-3
-2
-1
0
1
2
3


ICC for item v16
Probability
Theta
0
1
0.0
0.2
0.4
0.6
0.8
1.0
-3
-2
-1
0
1
2
3


ICC for item v17
Probability
Theta
0
1
0.0
0.2
0.4
0.6
0.8
1.0
-3
-2
-1
0
1
2
3


ICC for item v18
Probability
Theta
0
1
0.0
0.2
0.4
0.6
0.8
1.0
-3
-2
-1
0
1
2
3


ICC for item v19
Probability
Theta
0
1
0.0
0.2
0.4
0.6
0.8
1.0
-3
-2
-1
0
1
2
3

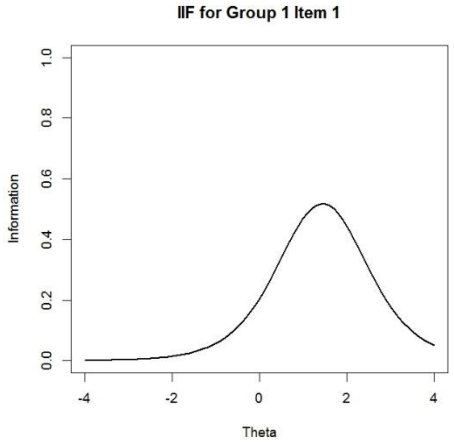
APPENDIX D

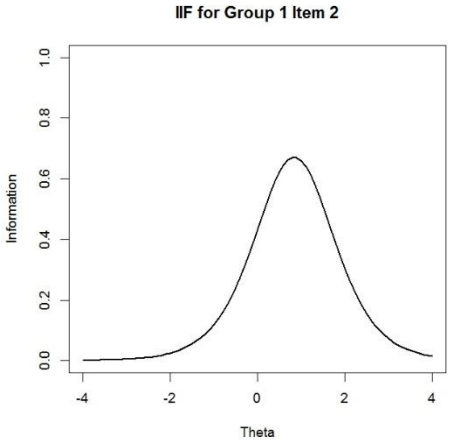
ITEM FIT STATISTICS

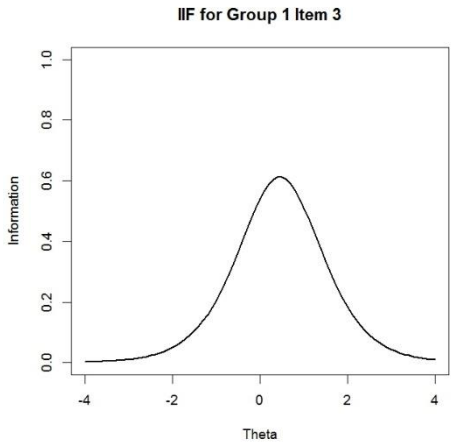
Item	S-χ^2	*p*
1	9.1	0.94
2	26.1	0.07
3	16.2	0.44
4	22.6	0.12
5	15.2	0.58
6	9.5	0.92
7	39.1	0.00
8	14.6	0.56
9	7.9	0.95
10	19.5	0.24
11	18.2	0.38
12	12.1	0.80
13	14.5	0.64
14	14.6	0.62
15	32.2	0.01
16	18.6	0.35
17	14.8	0.61
18	38.1	0.00
19	17.1	0.45

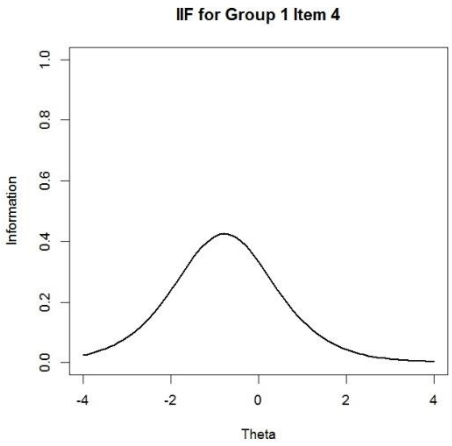
APPENDIX E

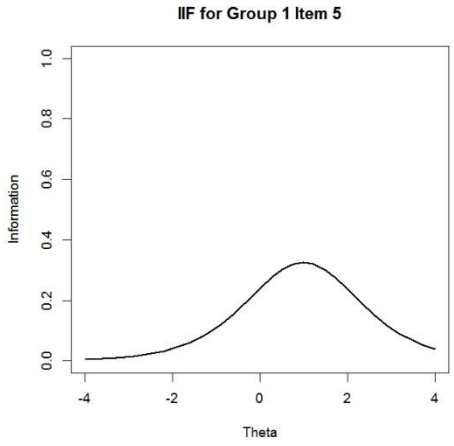
ITEM INFORMATION FUNCTIONS

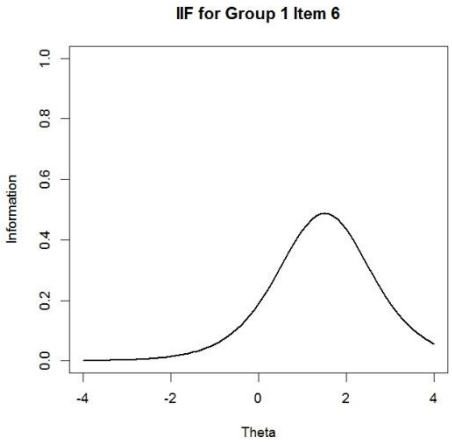


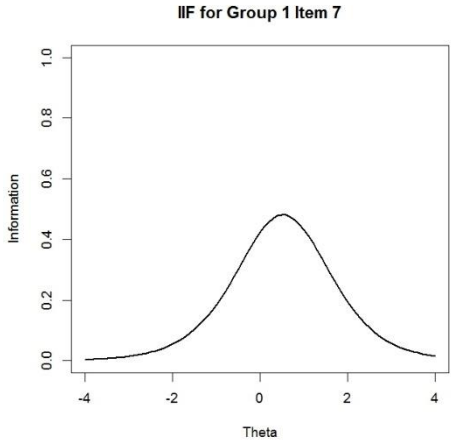


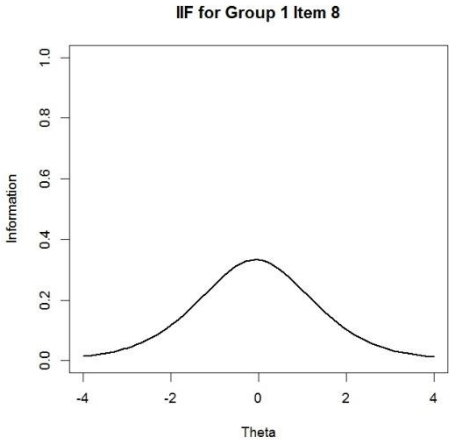


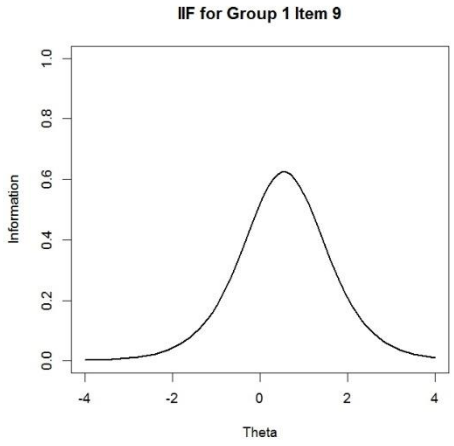


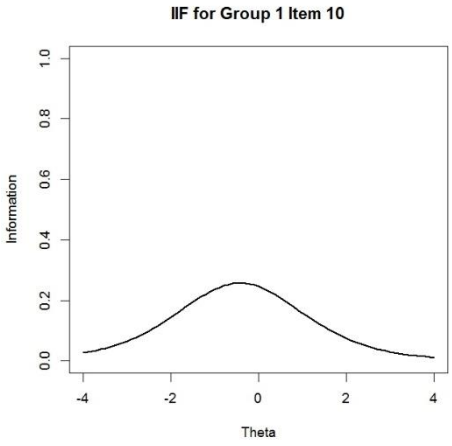

IIF for Group 1 Item 5
Information
0.0
0.2
0.4
0.6
0.8
1.0
-4
-2
0
2
4
Theta

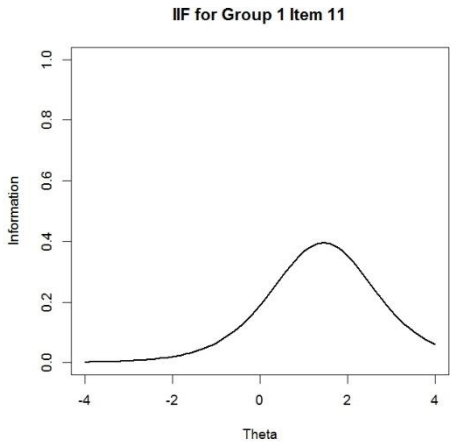

IIF for Group 1 Item 6
Information
0.0
0.2
0.4
0.6
0.8
1.0
-4
-2
0
2
4
Theta

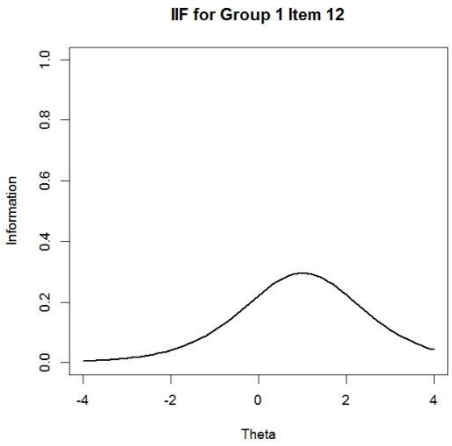

IIF for Group 1 Item 7
Information
0.0
0.2
0.4
0.6
0.8
1.0
-4
-2
0
2
4
Theta

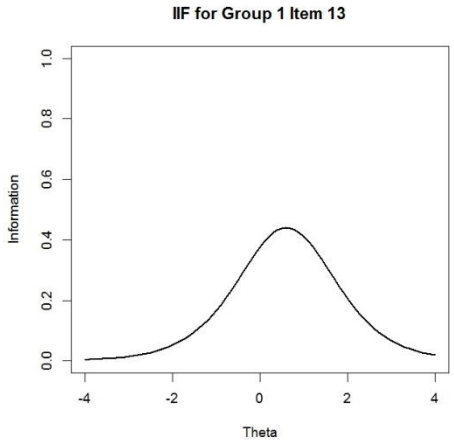

IIF for Group 1 Item 8
Information
0.0
0.2
0.4
0.6
0.8
1.0
-4
-2
0
2
4
Theta

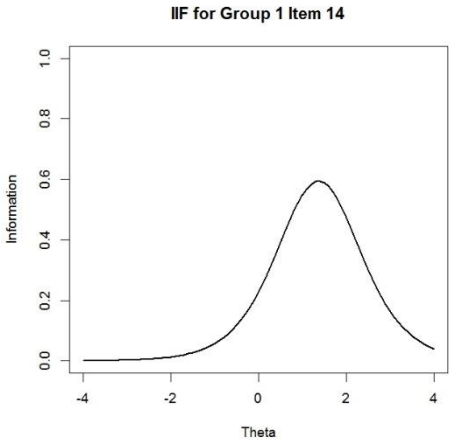

IIF for Group 1 Item 9
Information
0.0
0.2
0.4
0.6
0.8
1.0
-4
-2
0
2
4
Theta

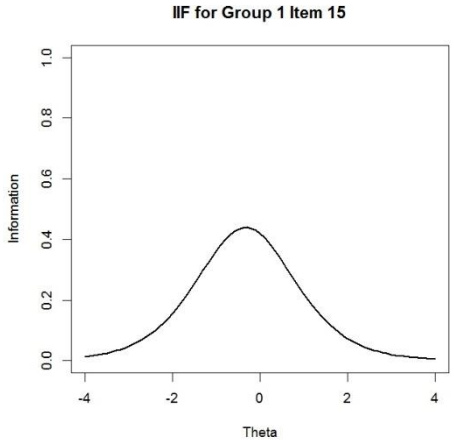

IIF for Group 1 Item 10
Information
0.0
0.2
0.4
0.6
0.8
1.0
-4
-2
0
2
4
Theta

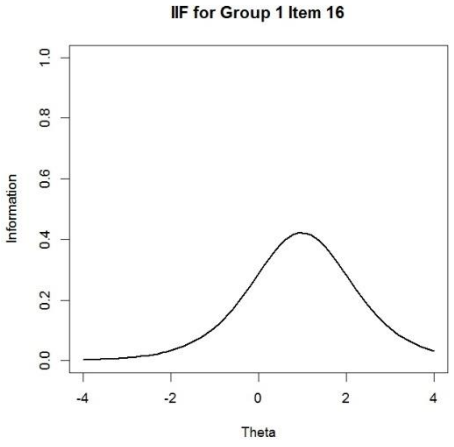

IIF for Group 1 Item 11
Information
0.0
0.2
0.4
0.6
0.8
1.0
-4
-2
0
2
4
Theta

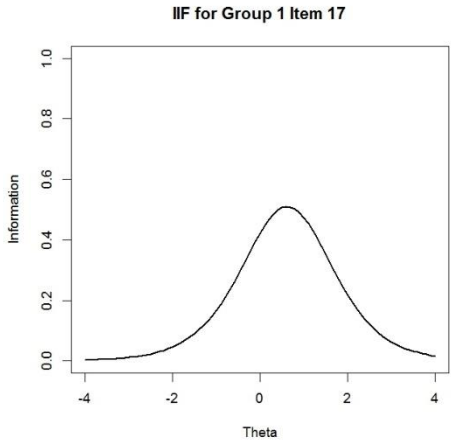

IIF for Group 1 Item 12
Information
0.0
0.2
0.4
0.6
0.8
1.0
-4
-2
0
2
4
Theta

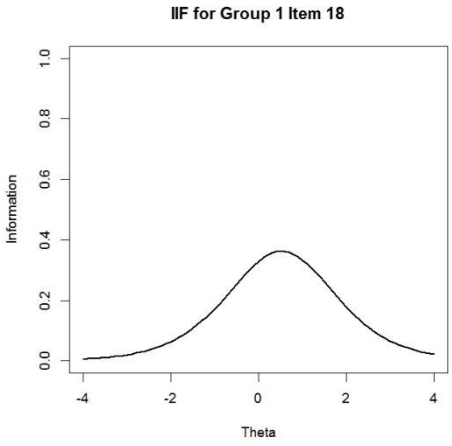

IIF for Group 1 Item 13
Information
0.0
0.2
0.4
0.6
0.8
1.0
-4
-2
0
2
4
Theta

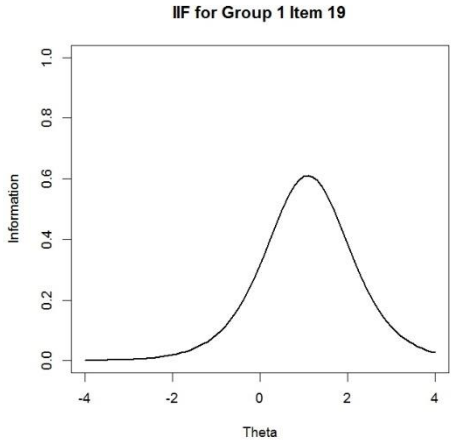

IIF for Group 1 Item 14
Information
0.0
0.2
0.4
0.6
0.8
1.0
-4
-2
0
2
4
Theta


IIF for Group 1 Item 15
Information
0.0
0.2
0.4
0.6
0.8
1.0
-4
-2
0
2
4
Theta


IIF for Group 1 Item 16
Information
0.0
0.2
0.4
0.6
0.8
1.0
-4
-2
0
2
4
Theta


IIF for Group 1 Item 17
Information
0.0
0.2
0.4
0.6
0.8
1.0
-4
-2
0
2
4
Theta


IIF for Group 1 Item 18
Information
0.0
0.2
0.4
0.6
0.8
1.0
-4
-2
0
2
4
Theta


IIF for Group 1 Item 19
Information
0.0
0.2
0.4
0.6
0.8
1.0
-4
-2
0
2
4
Theta